

**An Introduction to**

**PROGRAM  
EVALUATION**

**For your social service agency**



## INTRODUCTION

---

This manual is designed for individuals who work in agencies that help people with a variety of social service needs, such as public health, substance abuse, youth crime and violence, early childhood, after school programs, and mentoring services. Given the move towards evidence-based practice in our country, you are probably either already engaging in some level of evaluation activities or contemplating doing so. For better or for worse, you likely recognize that it is no longer good enough to have a *belief* that your agency helps people; nowadays, you must show *evidence*. This manual was designed to help you begin or improve this often daunting process of demonstrating the effectiveness of your program.

It can seem that evaluation is too far removed from your clients and your program to be valuable for you and your agency. In a world of limited funding, policymakers now have to choose among agencies, programs, and practices, all of which are intended to help people. This manual adopts the philosophy that, although evaluation may seem like something you have to do only to stay afloat, there are many ways that you, and your evaluators – whether internal or external – can make evaluation activities and results useful. We hope that this manual will help you feel invested in the evaluation process and, thus, help make it something that will ultimately improve your agency and, consequently, the lives of those you serve.

This manual also adopts the philosophy that at some level, all program evaluations involve *partnerships* between program professionals and evaluation professionals, and that neither one can do their job without the other. This does not mean that an evaluator must be external, but it does mean that, even if internal, an evaluator must be a person who has research methods and data analysis as the focus of their training and expertise.

This manual provides you with information on the following aspects of evaluation: fears and myths of evaluation, evaluation components, logic models, measure selection, survey administration, participant confidentiality, and data results and interpretation. In addition, the appendices include additional information intended to be helpful for evaluation activities, such as terms/definitions and sample consent forms. We hope that you will read and share this manual with staff at your agency and use it as a starting point for a discussion on how to add to, improve, or formalize your evaluation activities.

## TABLE OF CONTENTS

---

Chapter 1 Program Evaluation: The Big Picture .....	4
Fears and Realities of Program Evaluation.....	6
Chapter 2 Components of Evaluation .....	8
Chapter 3 Logic Models: The Evaluation Building Blocks.....	12
Chapter 4 Measure Selection.....	15
Measure Selection FAQs .....	15
Chapter 5 Survey Administration and Confidentiality .....	20
The Use and Importance of Unique Identifiers.....	21
Confidentiality, Respectful Survey Conditions, and Participants’ Rights .....	22
Will a Consent Form be Necessary?.....	22
Confidentiality.....	23
Final Checklist for Respectful and Confidential Survey Procedures .....	23
Chapter 6 Data Results: “Statistics for Dummies” .....	25
Statistical Significance or Large Effect Sizes.....	26
No Statistical Significance or Small Effect Sizes .....	27
Conclusion .....	29
References .....	30
APPENDICES .....	31
APPENDIX A: Evaluation FAQs.....	32
APPENDIX B: Sample Informed Consent Form.....	36
APPENDIX C: Sample Survey Instructions .....	37

## CHAPTER 1

### PROGRAM EVALUATION: THE BIG PICTURE

---

“Evaluation is the systematic application of scientific methods to assess the design, implementation, improvement or outcomes of a program.”

- Rossi and Freeman (1993)

The key word in the previous quote is *systematic*. In evaluation terms, the opposite of systematic evidence is when we have beliefs, perceptions and ideas about the effectiveness of our programs, and this is referred to as *anecdotal* evidence (as in anecdote, or story). Anecdotal evidence is important and contributes to the day-to-day morale and functioning of staff and clients within programs. Program staff can not likely be effective in their jobs if they do not *believe* that their program works. Systematic/scientific evidence is also indispensable, however, because it is more easily communicated to and understood by people external to your program, but who still have a stake in it.

All the people who have a stake – emotional, financial, or otherwise - in a program are referred to as *stakeholders*.

Thus, the goal is not to replace anecdotal evidence with systematic/scientific evidence, but rather to recognize the importance and proper place of each.

Anecdotal evidence of success = *Internal* functionality of your program.  
Systematic/scientific evidence of success = *External* sustainability of your program.

There was a time when anecdotal evidence was sufficient. After all, shouldn't it be obvious that people in need of services are better off with some services than none at all? In an age of increasing scientific awareness and capacity and decreasing funds for social programs, policymakers are in the position of having to pick and choose among programs – all of which have good intentions, and all of which are providing services for people in need.

For better or for worse, external stakeholders will support programs based on systematic/scientific evidence showing which programs are actually having an impact on people's behavior and improving their lives in measurable ways.

Anecdotal evidence can also be useful in choosing your *approach* for obtaining systematic/scientific evidence. A clear, personal understanding of how your program is helpful to people may help you choose among countless possibilities of how to systematically measure your program. To begin putting real concepts into the evaluation big picture, please read the following description of a program:

### Program Example

You run a comprehensive case management program for low-income, single mothers. Your program uses a variety of methods to help these women “get their lives back on track,” “become more connected to the resources in their communities,” and perhaps “become strong self-advocates who can get help when needed for themselves and their children”.

True, these are very broad goals, but they capture your program accurately because different women come to you with different strengths and needs. Consequently, you may feel that there is no one measurable outcome that will apply to all your clients.

When you take a moment to ask yourself, “How do our services help our clients?” you conjure up mental pictures of the different women on your caseload sitting in front of your desk, talking about their lives and their goals, and then you remember: You have spoken with many of them about going on hormonal birth control shots and/or their feelings about waiting a while before having more children.

Initially, delayed pregnancy may not seem like a compelling measurable index of your program’s success because first, it isn’t a direct equivalent of “getting your life back on track”, and second, it is only measured by time passing, with nothing happening! Why would *that* be convincing to an external stakeholder?

You then find research demonstrating that pregnancies within 6 months of a previous one are at 40% increased risk for premature birth<sup>1</sup>, which is associated with a host of medical and developmental problems that are costly and time-consuming to treat. In addition, you find that delayed pregnancy is associated with women being in a better position to complete school, obtain employment, leave welfare, and provide more positive child-rearing environments for their children<sup>2</sup>. You begin to find ways of systematically documenting your clients’ patterns of birth control use and pregnancy rates at their first contact with your program, by the end of the program, and after the program ends.

This example was intended to demonstrate that a seemingly “small” outcome can actually provide very compelling evidence of the effectiveness of your program. “Small” outcomes, or *proximal outcomes*, are most convincing when you can document their relationship to larger-scale/longer-term outcomes, or *distal outcomes*.

***You don’t have to,  
cannot,  
and should not try to:***

**Prove that your program “fixes” all problems,  
for everyone, all the time.**

In sum, you will not be held to an impossible standard for demonstrating effectiveness, and you can begin by thinking about seemingly small impacts that your program has on your clients.

## Fears and Realities of Program Evaluation\*

At this point, you may already be thinking about how to evaluate your program, or how the program evaluation you already have in place fits in to the definitions and goals we've discussed. There tends to be a host of preconceived notions about program evaluation – some valid and some invalid or at least outdated – so let's take this time to address some fears and realities of program evaluation.

**Fear:** *Program evaluation is merely something I am obligated to do to assure continued funding. Because it is all about “the bottom line”, staff will feel resentful and constantly under the pressure of “fighting for their life”.*

**Reality:** Some cynicism and over-focus on the financial implications of program evaluation is understandable in light of decreasing funds for social programs. However, cliché as it sounds, you get out of evaluation what you put into it. Upper-level staff in particular have a responsibility to “own” the idea of evaluation as an important and routine part of running a program. Staff generally do feel invested in improving their programs and serving their clients better, and if they understand that the best way to achieve this is through systematic evaluation, their ownership of the process will also increase. Although the more obligatory aspects of evaluation cannot be eliminated completely, they become easier to accept and carry out the more that staff can see that the results of evaluation are directly useful to them.

**Fear:** *Evaluation is a useless activity that generates lots of boring data with useless conclusions. I am willing to see it as useful to me – it's just that it isn't.*

**Reality:** This was a problem in the past when it was thought that program evaluation methods needed to mimic “perfect science”. The findings in reports oftentimes reiterated the obvious, and conclusions and recommendations were avoided. Understandably, this left program staff disappointed and skeptical that evaluation could ever be useful. Since that time, the field of program evaluation has worked hard to do its part in making evaluation useful to internal as well as external stakeholders. *Utilization-focused evaluations*, or evaluations for the purpose of “continuous improvement” are the norm now, and programs no longer have to worry that they will be asked to hold themselves to a standard that even the “state of the art” cannot meet.

**Fear:** *Evaluation is a highly unique and complex process that occurs at a certain time in a certain way, and almost always includes the use of outside experts. Since I'm not a researcher or statistician, my choices are to conduct an evaluation that won't have credibility or to call in an outsider who doesn't understand my program, or who will provide results I don't understand.*

**Reality:** Program staff and supervisors do not need to perfectly understand scientific terminology to conduct or understand high-quality evaluations. They should use their expert knowledge about their program to think about what information they will need to make the best decisions for their program. In evaluation, form follows function. The specific methodological choices will be driven by your conclusions about the information you need to maintain and improve your program. In cases where outside experts are needed, it is their responsibility to make the results understandable and useful to you. Do not make contracts or partnerships with any evaluation experts unless they provide you with a specific outline of how the results of the evaluation will be made accessible to you. Conversely, do not call in an outside expert to “take care of everything” and think that you

---

\* Adapted from McNamara (2006).

have met your evaluation obligations. You must work collaboratively with evaluators to get the most out of the process.

**Fear:** *Program evaluation is about proving the success or failure of a program.*

**Reality:** Research of any type rarely provides simple “yes or no” answers, nor does it seek to do so. Most evaluation reports will describe some findings in the desirable direction, some in the undesirable direction, and some *null* findings (results that show no effect either way). Programs rarely, if ever, live or die based on the findings in one evaluation report. All stakeholders have the responsibility to interpret findings in context. The best interpretations are those that go beyond statistical significance to inform decisions about program improvements.

Conversely, it is not realistic to think that success means that the program has been implemented perfectly and that you will never have to hear from employees, clients, or funders again. This doesn’t happen in real life, nor is it the best use of evaluation, which needs to be an ongoing, or *iterative* process. Iterative means that modifications to a program are made in increments with evaluation in between each step to continuously inform the next iteration.

### Chapter 1: Take-Home Points

- Anecdotal evidence is important for the internal functioning of programs because it shows that program staff believe in what they are doing.
- Systematic/scientific evidence requires more effort to obtain, but is indispensable because it explains your program to stakeholders who don’t have the advantage of seeing the program function day-to-day.
- Program evaluation uses systematic/scientific methods to get information about programs, but your in-depth understanding of the program will be used to inform the methods and approach of the evaluation.
- Your program cannot be everything to everyone and that is not the standard to which you will be held. Think about proximal outcomes: smaller rather than larger, specific rather than general, and closely tied to your program’s goals.
- Although parts of the evaluation process aren’t always fun, a belief that it was worthwhile will be based on its ultimate utility, or usefulness, in continuous improvement of your program. You, your evaluators (whether internal or external), your funders, policymakers, and social scientists in the field of program evaluation at large all have a responsibility in maximizing the usefulness of evaluation processes and products.
- If you are just beginning the process, don’t begin by worrying about what type of evaluation you need to do. Start by asking yourself, “What do I need to know to make the best decisions for my program?” The specifics of the evaluation will follow from the answers to that question.

Most importantly, evaluation should be thought of as a *standard* part of running a program. When program plans are being made (budgets, hiring, goals, scopes of work, etc.), evaluation should be factored in. The more all staff can adopt the attitude of, “How can we function well if we don’t know how well we’re functioning?”, the more evaluation will seem a natural part of the process, with obvious benefits.

## CHAPTER 2

### COMPONENTS OF EVALUATION

---

The field of evaluation science has many different categorizations of evaluations that can be done (e.g., needs assessment, cost-benefit analysis, effectiveness, efficiency, summative, formative). The most basic two types of program evaluation are *process evaluation* and *outcome evaluation*.

Process evaluation describes the “who,” “what” and “how” of program services. It answers the question, “Is the program being implemented as expected?”

Outcome evaluation describes behaviors, or changes in behaviors of clients as a result of the program. It answers the question, “What was the impact of the program on the clients served?”

Your interest may have focused quickly on the second type – outcome evaluation – because after all, isn’t this what we really want to know? If we prove positive program impacts can’t we assume (or disregard) that the program was implemented as intended?

YES!

Did you think the answer was going to be no? The answer to the question is yes, but as always, the devil is in the details.

*Proving that your program caused a positive impact* on people is the ideal we strive for, but near-impossible to achieve in reality.

There are very few techniques that are widely accepted in science as adequate proof of causality. Usually, researchers who want to make strong claims about one thing causing another need to employ *randomized controlled trials* (RCTs), which are very expensive and very time-consuming. Moreover, unlike more traditional RCTs, such as those used to test the effectiveness of medicines (with placebo conditions), social science RCTs cannot be as tightly controlled (e.g., cannot have placebo conditions where people *don’t know* if they are in the treatment or control group). Thus, even the best available methods in social science do not result in our ability to claim causality.

Thus, a process evaluation cannot be bypassed because:

- 1) If you *do not* show positive impacts in your outcome evaluation, the process evaluation can help you figure out *why*.
- 2) If you *do* show positive impacts in your outcome evaluation, the process evaluation can lend support to the idea that your program was responsible for those results.

Think about how the following information could lend credibility to good results in your outcome evaluation:

- Client “no-shows” were reduced from 24% one year to 16% the following year.
- The more services an individual received, the stronger were the positive results – this is known as *dosage* effects.
- The more clients reported liking their caseworkers, the stronger were the positive results.
- Staff turnover was less than average this year.

All of this information about services, client perceptions, and implementation of your program is derived from the process evaluation! While we still cannot prove causality, we have a richer picture of resources in our “information arsenal” to help stakeholders put the evaluation in context. Undesirable or null findings are not interpreted in isolation, and desirable effects can be more plausibly attributed to your program.

So now that you’re convinced of the importance of process and outcome evaluation, let’s discuss some possible components of each:

#### Ideas for a Process Evaluation

You already keep records about your program and services. This comprises a significant part of the process evaluation, but it may involve some formalizing or organizing of the different elements. Some examples of process information to track are:

- Number of staff and their background education and training
- Demographic information (e.g., age, ethnicity, gender, languages spoken) about clients and staff
- Number of *planned* activities/services and number of *actual* activities/services
- Length of staff’s employment with the program and turnover rates
- Client attendance
- Goals for program growth and milestones toward those goals

A really rigorous process evaluation includes not only facts and figures about how your program is being carried out but also information about the quality of services you provide. Examples of qualitative program information include:

- Client satisfaction surveys including their perceptions of the staff with whom they work
- Objective observations of “live” services
  - Attending a session of a parenting course with a specific curriculum to observe adherence to the curriculum
  - Observing an after-school program to assess tutors for using both nurturance and structure with students
  - Shadowing a home visitor in a post-partum home visiting program
- Focus groups with staff to ask about support they need to carry out their jobs better
- Assessment of procedures in place for training new staff and continuing education opportunities for staff

These are simple concepts, but tracking them *systematically* requires planning and organization. Discuss with your staff ways you can build upon tracking methods you already have in place to capture additional information that will help demonstrate whether your program is being implemented as intended.

## Ideas for an Outcome Evaluation

A typical outcome evaluation of small-to-moderate scope will usually have one of two approaches:

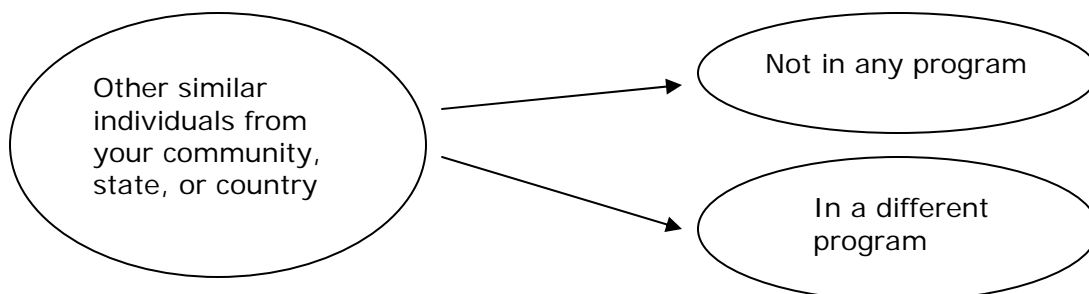
1. A targeted goal for the frequency of a behavior.
  - 85% of youth will report no drug or alcohol use within the past 30 days by the end of the program
  - 90% of women will not become pregnant within a year
  - 75% of youth will score in the "Proficient" or "Exceptional" range on the assessment
2. A pre/post design in which the goal is statistically significant improvements in the targeted behaviors.
  - Significant increase in school bonding/attachment
  - Significant decrease in family conflict
  - Significant increase in reading skills
  - Significant decrease in tobacco use

In the case of Approach 1, credibility of the chosen goal, and therefore the results if that goal is met, will be increased if it is based on already-known population or community statistics about the same issue, also known as *baseline statistics* or a *baseline comparison*. While stakeholders have some responsibility to research this information as well, the reality is they often don't. As a result it may be easy to underestimate a finding that doesn't *seem* important enough (e.g., 12% of youth quit smoking, but it turns out that this is a 1200% improvement over the only 1% who quit *without* a cessation program).

In the case of Approach 2, there is the advantage of statistical significance providing an internal benchmark of, "Is this amount of change a big enough change?" However, statistical significance is negatively affected by sample size and programs usually have relatively small samples. Thus, any baseline data about change that would be expected without a program or with a different program, will boost the credibility of this approach as well. In rare cases you may have your own comparison data available, perhaps from a previous version of your program, or from another school not implementing your program.

Any comparison data that answers:

How the clients in your program changed relative to:



Lends strength to and places in context the results of your evaluation.

A comparison group is not always available; however, evaluation can still benefit your program. Outcome evaluation is very useful for program improvement, but needs to be well planned and organized before implemented. Logic models (discussed in Chapter 3) assist in the process of planning and organizing program evaluation.

### Chapter 2: Take-Home Points

- Process evaluation and outcome evaluation are complementary in nature.
- Process evaluation describes the “who,” “what” and “how” of program services. It answers the question, “Is the program being implemented as expected?” It can help explain why you got your results.
- Outcome evaluation describes behaviors, or changes in behaviors of clients as a result of the program. It answers the question, “What was the impact of the program on the clients served?”
- Proving that your program caused a positive impact on people is the ideal we strive for, but near-impossible to achieve in reality. Causality cannot be claimed through evaluation.
- Despite the fact that causality cannot be proven in program evaluation, there are ways to boost the credibility of your results.
  - Improve the level of detail and systematization in how you collect data about your program.
  - Do your homework about baseline statistics for your targeted outcomes in your community, state, and the national level.
  - Obtain data about how people improve or not when they are A) not in any program and B) in a different program type than yours.

## CHAPTER 3

### LOGIC MODELS: THE EVALUATION BUILDING BLOCKS

---

For purposes of this chapter, we will assume that we are operating within your already existing program. That is, you work in a program that already has a target population and one or more social issues you are trying to target. You are beyond the stage of just looking for ways to spend a grant, start a program, or do a *needs assessment*.

Yet, even the longest-existing and best-run programs don't always have a very concrete idea of what they do, whom they serve, and how they accomplish their goals. These processes are often implicit rather than explicit.

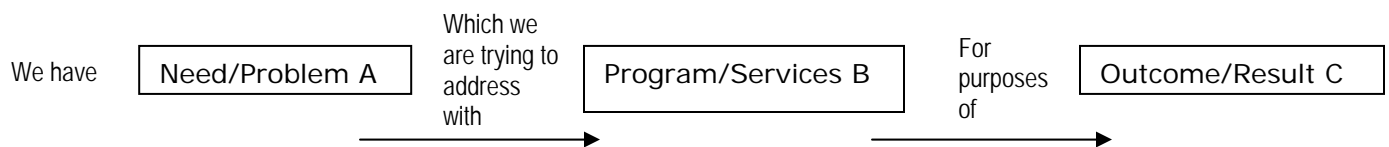
The logic model is a handy tool that:

- formalizes and organizes your goals for internal stakeholders
- explains and defends your program to external stakeholders

Logic models don't have anything to do with statistics or even complicated research methods. They simply outline a *logic* that ensures that individuals coming from different contexts, understandings, and organizations will understand the basic program activities, goals and assumptions, and thus be on the same page. The different stakeholders don't necessarily need to agree on everything in the logic model, and the logic model may contain assumptions that are false (e.g., pie-eating program will reduce obesity). The logic model forces us to “put it all out on the table” for public viewing and assessment. This aids in the iterative process of continuing to question our assumptions, tighten up our logic, and ultimately improve programs.

If you can create an “IF...THEN” statement, you can create a logic model.

There is no single correct logic model. They can range from simple to complex. The most basic sequence with respect to “ABC Social Program” is:

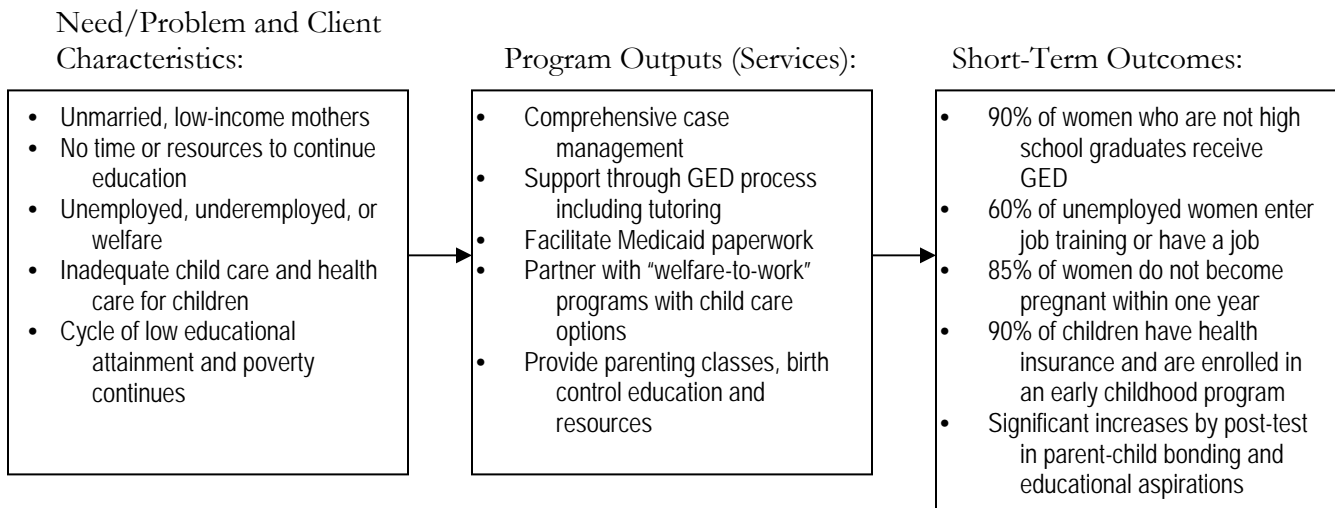


Again, it may seem obvious how your program fits into this simple model, but perhaps some elements are clearer than others. If you recall our example of a comprehensive case management program for low-income mothers, the “need/problem” isn't immediately clear. Comprehensive programs tend to be more successful on the one hand because they treat the “whole person”, but on the other hand, it may be more difficult to specify one or a few issues that are the top priorities. Conversely, you cannot necessarily trace back in an obvious way what the need/problem is by knowing what the outcome/result is. Following the example, if “delaying subsequent pregnancies” is the targeted outcome, this doesn't necessarily mean that “too many pregnancies close together” was the initial need/problem. Your program may need some “interim steps” in the logic model in

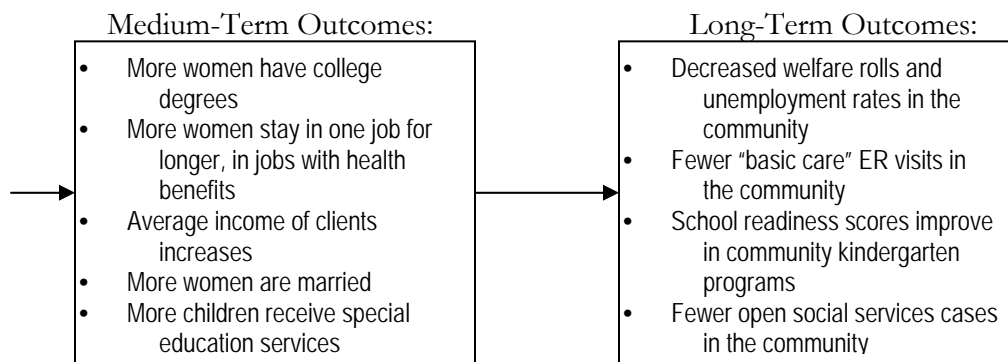
order to justify the links. For the case study example, we would need to logically connect too many pregnancies close together to inadequate opportunities for education and self-sufficiency. For example, **IF** women have time and financial resources not being spent on an additional child, **THEN** they will have additional opportunities to obtain an education and financial stability.

Programs will be “judged” based on short-term, proximal outcomes. However, your logic model and therefore your “arsenal of information” will be still strengthened if you can document the links between your outcomes and more distal ones.

A simplified logic model for our comprehensive case management program might look like this:



Even though your program is not responsible for medium-term and long-term outcomes, creating logical links to those more distal outcomes spells out your mission to external stakeholders, and gives credence to the importance of your proximal outcomes. So, the next boxes in your logic model might look like this:



Some stakeholders might expect, or you might want, your linkages to be even more specific. In the above example, elements in the logic model are grouped together, rather than linked one at a time. For example, you might need to specify that parenting classes are what specifically should lead to an increase in parent-child bonding, or that partnerships with welfare-to-work programs are what specifically lead to an increase in employment or participation in job training programs.

### Chapter 3: Take-Home Points

- A logic model is beneficial for both internal and external stakeholders. It formalizes and organizes your goals for internal stakeholders and explains and defends your program to external stakeholders
- IF-THEN statements can assist you in developing your logic model by helping to clarify and articulate the need or problem your program is attempting to alleviate.
- Logic models are a starting point to illustrate to stakeholders how your program's services are connected to the expected outcomes.
- Short-term or proximal outcomes are the ones that you are expected to measure at the end of clients' time spent in your program services.
- Including medium-term and long-term outcomes in your logic model is also helpful to inform external stakeholders of your overarching mission.

## CHAPTER 4

### MEASURE SELECTION

---

The next level of specificity in the logic models we discussed in the last chapter involves the actual measure, survey, tool, or instrument that you will use to measure your outcomes. If you are using descriptive benchmarks such as a certain percentage who will obtain a GED, a simple count of GEDs may be all that is required. But some benchmark outcomes will still need to be defined, or *operationalized*, with a greater level of specificity. Delayed pregnancy, for example, is not an operationalized construct. We would have to specify whether we meant any known pregnancy at all or only those that resulted in a live birth, and we would also have to specify the minimum duration of time that would be considered a delay (e.g., no pregnancies beyond 12 weeks gestation for at least two years).

For pre- and post-testing, you will likely need to select a published measure that has documented *validity* (the measure actually measures what it says it does) and *reliability* (the measure demonstrates reasonably consistent results across time, settings, etc.). Of course, you want a measure that applies to and fits with your program and your logic model. As with all evaluation issues, measure selection is an issue of balance.

You should select a measure that fits your program and your logic model, but one that isn't so specific that it can't be generalized beyond your program.

Just as you will work with your evaluators (whether internal or external) to develop your logic model, you will also work with them to select a measure that validly and reliably captures the short-term, proximal outcomes specified in your model.

#### Measure Selection FAQs....

##### **1) How do I find a measure that accurately captures my program's proximal outcomes specified in my logic model?**

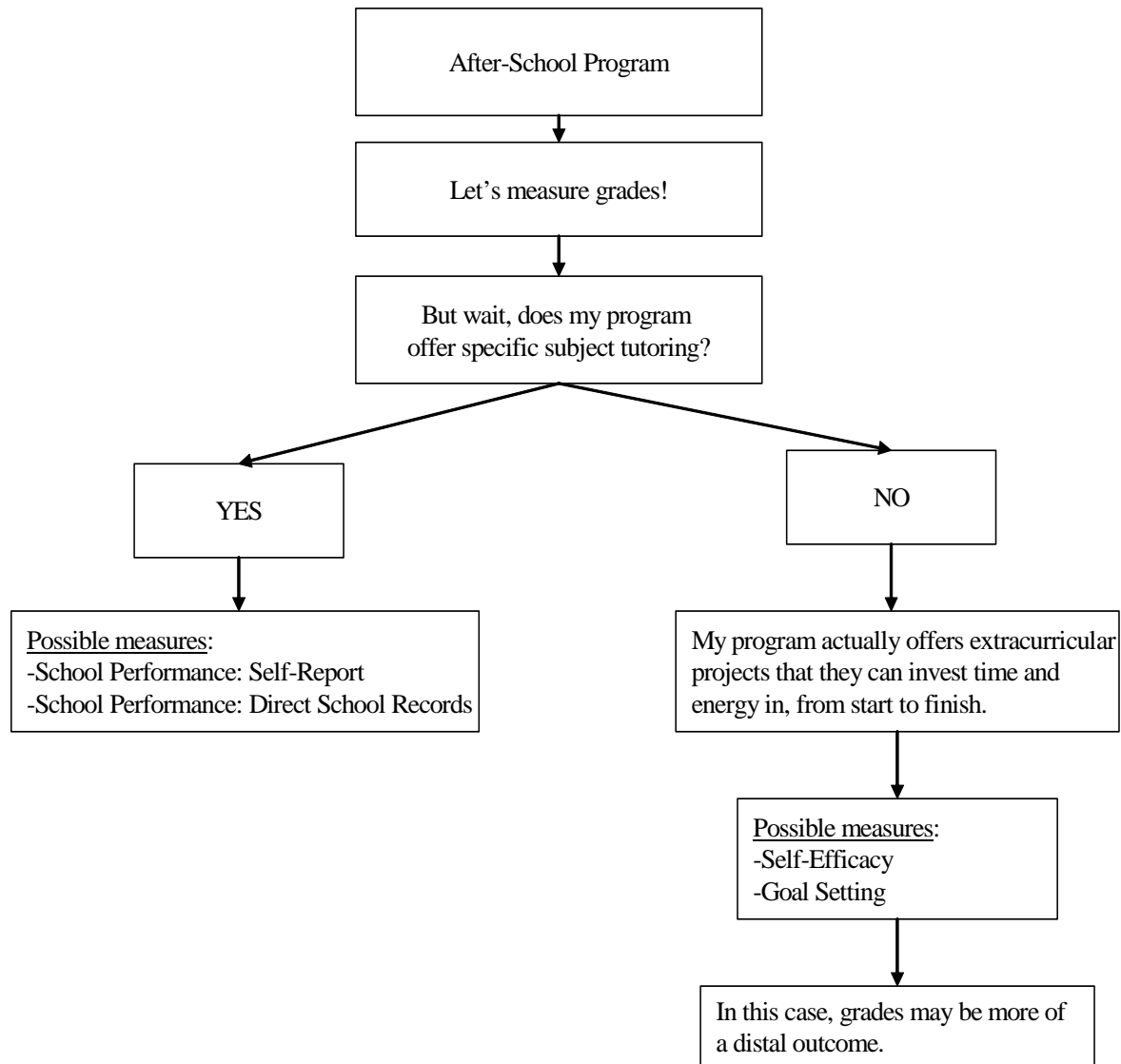
If your measure is not pre-selected or chosen from a permissible list, there are various avenues of searching for, finding, and finally choosing a measure that will assess your program goals and expected outcomes as indicated in your logic model. Evaluators typically have knowledge of existing measures that are statistically valid and reliable. Working with your evaluators will help make the decision as to the best measure to assess your outcomes.

When the assistance of an internal or external evaluator is not available, internet search engines and library databases are an obvious choice. Be cautious in using generic internet search engines, as it can be difficult to verify whether a measure you find is published, valid, or reliable. A good rule of thumb is to check for a "Technical Manual" associated with the measure. Chances

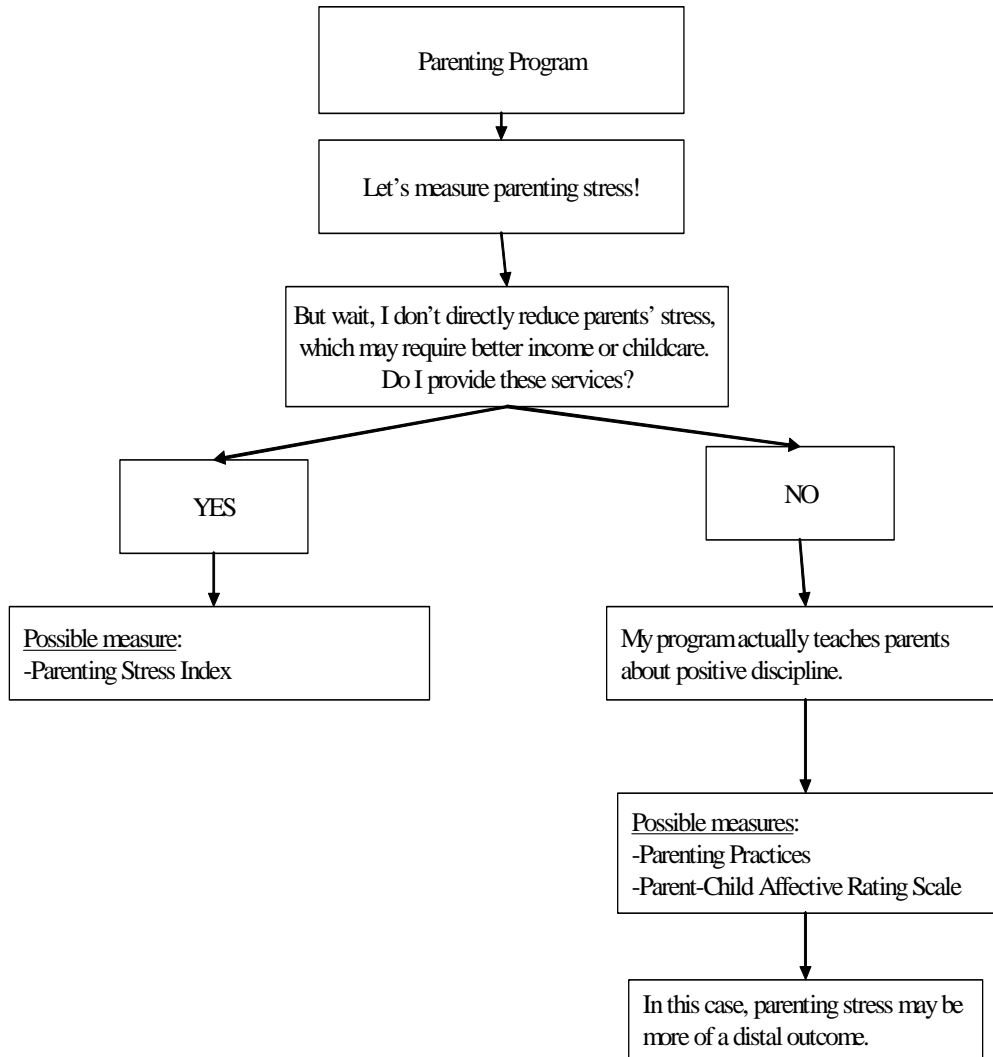
are good that a measure that has a technical manual has been subjected to scientific testing at the level of rigor that would be expected for program evaluation. A measure that is published in a scholarly, peer-reviewed journal also has a higher probability of having been adequately researched. Some measures are “in the public domain”, meaning that there is no fee or permission required to use them. Other measures require permission from the author or publishing company. Many times, you can obtain permission to use the measure for free or a very small amount.

Measure selection is a step-by-step process that needs to be thoroughly and logically planned, drawing connections between expected outcomes and what the measure is asking. The following “decision trees” are examples of programs and the process they might go through to choose a measure:

**EXAMPLE 1:**



## EXAMPLE 2:



As you can see, measure selection isn't always obvious from the type of program you administer. That is why you will want to have a good logic model specified before you try to choose a measure. The most important issue to consider is whether the specific services you provide, and the length of time in which you provide them, can logically and realistically be expected to impact the items assessed in your chosen measure.

### 2) What if I don't have a choice about which measure to use?

In some cases your agency may be taking part in a larger-scale evaluation of services and as a result you may be required to give a certain measure, or may have to choose from a list, which can feel restrictive.

Remember that there is always a trade-off between breadth and depth: While a required measure may not fit your program perfectly, the broader knowledge gained from your data makes this a worthwhile compromise.

Though the benefits to your agency are indirect, it ultimately serves clients and programs better when evaluation is not restricted to agency-by-agency findings that cannot be generalized or interpreted as a whole. By participating in an evaluation beyond the scope of your single agency, you are helping advance evaluation science, which will ultimately contribute to the stability of prevention and intervention service delivery.

Also remember that your participation in a large evaluation can be only one part of your evaluation efforts. You can and should choose measures beyond the scope of a required evaluation that answer the questions you want, in the manner you want. As always, choices will have to be made carefully and in consideration of your logic model. Furthermore, you will want to seek balance between potential knowledge gained and the burden on your clients. Generally speaking, asking your clients to fill out one or two additional surveys (depending on their length) is not perceived as undue burden.

### **3) Why are evidence-based measures negatively phrased or “deficit-based”? My clients feel judged and as a result I have difficulty obtaining their participation.**

It is true that no matter how often we explain to clients that the measures assess “programs” and not “people”, it is inevitable for an individual answering a survey to feel that they have a burden to answer in the “right” way. It can feel difficult as program managers to be cheerleaders for participating in research when we ourselves have ambivalent feelings about the measures, or don’t feel we can stand behind the approach of the measures.

No one wants to be in the position of having to apologize for evaluation research. Moreover, when we take the stance that evaluation is an unpleasant enterprise that an external force is “making” us do, it causes an “us vs. them” (i.e., providers vs. researchers) mindset, which further decreases clients’ investment in the evaluation process. That is why this manual keeps repeating the idea that all parties have ownership in the evaluation process, and a responsibility to promote evaluation as a necessary and natural part of providing services.

As far as measures are concerned, questions on surveys have to be worded in a variety of ways – positive, negative, and neutral – in order for the statistical analyses to take into account the fact that fallible human beings are the ones who fill out surveys. If we had surveys that were worded only positively, we would still have “human error”, but the statistics would be unable to account for it because there would be no basis for comparison. Features of measures that can seem annoying – like “deficit-based” wording, or repeating the same theme over and over – are actually necessary techniques that help the statistical analyses account for the fact that a measure of human perception will always be imperfect.

It is ok and even expected as part of the protocols for the protections of human subjects to show empathy for your clients around the discomfort that filling out measures can sometimes cause. At the same time, you don't want to send the message that you don't believe in the measure and that you are only handing it out because you have to.

Here is some language we can recommend to help you start finding this balance with your clients:

“Sometimes surveys ask tough questions, questions that have unpleasant content, or questions that seem to ‘expect the worst’. Please know that if some of the phrasing seems negative or judgmental, this is not reflective of how evaluators feel about this program or our clients. A variety of question types is necessary because it insures the most accurate results when it comes to compiling many people’s answers together.”

#### **4) Isn't it better to develop a measure that is specifically tailored to the goals and objectives of my program?**

There is an understandable tendency to want to create instruments from scratch, so that they can be as specific to one's program as possible. Despite the intuitive appeal of a customized instrument, this is generally not a good idea. Valid instruments take time and money to develop, test, and revise. Moreover, when an instrument is customized to a specific program, it starts to look more like anecdotal rather than scientific evidence because it cannot be generalized beyond the people or circumstances of your program. Once again, external stakeholders are interested in the “bigger picture” of how your program fits into the solution to social problems.

#### **Chapter 4: Take-Home Points**

- Remember that only proximal outcomes that can be measured by counts or pre/post surveys are the ones you will be expected to measure.
- If you have a required measure or list of measures due to participating in a large-scale evaluation, you will have the advantages of A) being able to compare your results to the larger context and B) having an easier job of measure selection.
- You should select a measure that fits your program and your logic model. Be sure that the measure isn't so specific that it can't be generalized beyond your program.
- A measure “from scratch” usually does not have sufficient credibility for external stakeholders. Selection of published measure will ensure validity and reliability.
- “Deficit-based” wording is a necessary technique used in survey development to account for human error. Help your clients understand that the question wording does not mean you or evaluators have negative expectations of them.
- Measure selection is an issue of balance: between breadth and depth when selecting the measure as well as a balance between gaining knowledge from participants and protecting the rights of participants.

## CHAPTER 5

### SURVEY ADMINISTRATION AND CONFIDENTIALITY

---

Once you know what measure(s) you are going to be using at your agency, you can begin putting procedures in place for handing out, collecting, and tracking the data. Your agency will modify the process to fit your situation, but there are certain “must-haves” that need to be in place. As repeatedly emphasized in this manual, program staff members will *set the tone* for participants regarding the value and importance of the survey effort, and can do this by ensuring proper survey conditions and conveying investment in the survey process. Program staff are also in the position to ensure that the data collection process is carried out in a way that will promote good data quality and protect participants’ confidentiality. The following procedures and tips are designed to ensure a successful survey effort.

**When collecting surveys, your two highest priorities should be:**  
- Ensuring the highest quality data possible  
- Protecting the rights and feelings of clients filling out the surveys

These priorities are interrelated. So let’s discuss the actual steps that need to be taken to meet these priorities.

When you acquire the survey(s) you will be using with your clients, you first need to make sure the format is user-friendly, and that it will capture all the information that you will need for purposes of the evaluation.

Generally speaking this means including on the “face page” of EACH INDIVIDUAL SURVEY:

- ✓ **A unique participant identification number for each person**
- ✓ General instructions for the survey
- ✓ An indication of whether it is pre-test, post-test, or follow-up
- ✓ A space for today’s date
- ✓ A space for your agency name or agency number if you have one
- ✓ A space for “cohort name” that includes the year and timeframe (e.g., “Fall 2006”, “March Visits, 2005”)

And often also included on the face page is:

- ✓ Participant’s birth date
- ✓ Participant’s gender
- ✓ Participant’s ethnicity
- ✓ Participant’s grade (if it is a child or adolescent survey)

**Participant names should not be on surveys to help protect confidentiality!**

## The Use and Importance of Unique Identifiers

The issue of unique participant identifiers is critical because:

1. It is absolutely essential if you are involved in a “pre/post” evaluation.
2. It involves more than just putting numbers on each survey.

Unique identifiers, if used properly, provide not only important confidentiality to your clients, but also a built-in tracking system, without which, an analyst cannot make any statements about improvements in your clients over time.

It is recommended that unique identifiers:

- ✓ Be four digits long
- ✓ Do not start with zero (allows for up to 8999 participants)
- ✓ Contain only numbers (no letters or other characters)

A master list of ID numbers, matching first and last names, and cohort name, will need to be maintained by one or two individuals at your agency. You can maintain a spreadsheet that is password protected, so only the individuals who need to be able to match people with IDs can be given access to the spreadsheet. Your data could look like the following spreadsheet:

Client ID#	Respondent Name	Fall 2006	Spring 2007
1000	John Smith	x	x
1001	Jane Doe	x	x
1002	Mary Brown	x	
1003	Sue Brown	x	
2000	Johnny Hanson		x
2001	Jamison Jay		x
2002	Stephanie Stills		x
2004	Martin Gove		x

The conventions for the above table indicate that you have pre- AND post-data for John Smith and Jane Doe, pre-data only for Mary Brown and Sue Brown (they may have dropped out of the program, not been present on survey administration day, etc.) and post-data only for the last four individuals. In this case, the individual doing the tracking decided to give “new” clients in the spring ID #s that started in the 2000’s rather than the 1000’s.

Generally speaking, you will provide one paper survey per person, *with the client ID already filled out* (do not ask participants to write or memorize their numbers). This means that you must already have created the master list that matches names and numbers prior to handing out the surveys to the participants. If you do not create this list before giving out the surveys, there will be no way to know what ID number to give the participants at post-test. It will not work to receive a group of surveys and number them afterwards because there is no way to know who filled out each survey, and again, no way to match the ID numbers at post-test.

Someone at your agency needs to maintain a confidential, password-protected master list of participant names and ID numbers. This list must be created before surveys are administered.

## Confidentiality, Respectful Survey Conditions, and Participants' Rights

The need for ethical standards to protect human subjects in research studies was brought to the forefront during World War II. In the name of medicine and science, Nazi doctors and scientists conducted cruel, harmful, and often deadly experiments on prisoners in war camps.

In 1946, after public awareness and uproar, leading members of the German medical hierarchy were indicted before the war crimes tribunal in Nuremberg, Germany. The trial resulted in the conviction of 16 of the 23 physician defendants, seven of whom were sentenced to death. The tribunal condemned the experiments and classified them as "crimes against humanity." "Permissible Medical Experiments," the standards used to judge these German doctors, became known as the Nuremberg Code and remain the basis for ethics codes internationally. For more information, please look at the following website: <http://ncjj.servehttp.com/irb/History.asp>

The principles of *informed consent* and *minimizing harm* are the focus of the Nuremberg Code.

Generally speaking, you will not need a formal ethics review for the survey procedures you conduct as part of your program; however, it is good practice to follow the federal and legal principles set forth for the ethical treatment of human subjects.

- Let your participants know their participation is encouraged, but voluntary.
- Inform them about what they will be asked to do.
- Make sure you are not unreasonably burdening them, and that the burden involved is worth the knowledge gained.

### Will a Consent Form be Necessary?

Because you will be engaged in evaluation research, whose purpose is to improve program services, a procedure called *passive consent* is usually acceptable. Passive consent involves informing participants (and their guardians in the case of minors) of the research to be conducted, and allowing the opportunity to decline participation. Thus, a passive consent form only needs to be handed in to you if someone does NOT want to participate – otherwise, consent is assumed. Please see the end of this manual for a sample passive consent form and a sample survey protocol.

It is strongly encouraged for you to include your passive consent forms in your initial registration packets, and introduce the idea of participating in the evaluation effort to clients upfront.

This is just one more piece in the puzzle of making program evaluation an integral part of service delivery at your agency.

## Confidentiality

One of the ethical and contractual requirements for all survey efforts conducted is to ensure that individuals' responses to survey items remain confidential. Some survey items are of a sensitive nature and it is especially critical that program staff, school administrators, faculty, community members, as well as other participants, do not have access to individual survey responses, whether accidental or intentional. Individuals must be assured that they can answer survey questions honestly without their responses or identities being revealed. Your evaluators should not provide individual-level data to any organization, but will report results in the aggregate.

Individuals' names should not appear on surveys and access to the password-protected master list should be limited to an as-necessary basis, and only for individuals who have knowledge of ethical treatment and confidentiality procedures.

Finally, and most importantly, individuals at your agency should not have access to client data once it is submitted\*, and should definitely not examine or keep their own compilation of survey responses. That is because evaluation surveys serve the purpose of evaluation research, *in the aggregate*, and are not intended to be tools for individual clinical intervention. Any attempt to match evaluation survey responses with an individual client is therefore a violation of confidentiality as well as a misuse of survey tools, which are not designed to help draw conclusions about individuals. If your agency is in need of formal, clinically-designed tools to aid in diagnosis or intervention, they should be tools designed for this purpose, and understood by all parties to be completely separate from evaluation surveys.

### Final Checklist for Respectful and Confidential Survey Procedures

- ✓ The program coordinator should meet with program staff and survey administrator prior to the survey administration to address confidentiality issues and review survey procedures.
- ✓ Program staff and survey administrator need to be made aware that it is their ethical, as well as legal, responsibility to avoid exposure to individual's responses (i.e., "no peeking", no access to master list without being authorized, for good reason and through the proper channels).
- ✓ Make sure participants also understand that they are not to put their names on any surveys they are asked to complete.
- ✓ Although participation is strongly encouraged, participants need to be made aware that they are not required to complete surveys and that they can skip questions they are uncomfortable answering.

---

\* Even when individuals at your agency are responsible for entering the survey data, there is no reason for them to have access to client names. That is why it is so important for client surveys to be identified only by a unique, numerical identifier. The one or two staff members who maintain the matched lists in order to *give out* the proper surveys should not be the same people who enter the data.

- ✓ The survey should be conducted in a quiet room with minimal distractions or disruptions. There should be enough space to allow participants some level of privacy when completing the survey.
- ✓ Survey completion will depend on the survey given. Allow for an appropriate amount of time. The survey administration process should not begin until all individuals are in the room and ready to begin. If a participant finishes early, send them out of the room to another activity, or provide them with quiet work until the survey time is up.

### Chapter 5: Take-Home Points

- When collecting surveys, your two highest priorities should be ensuring the highest quality data possible and protecting the rights and feelings of clients filling out the surveys.
- Rigorous survey administration utilizing unique identifiers, a participant tracking spreadsheet, and the management of this spreadsheet is crucial for matching participants during data analysis.
- This spreadsheet should be created prior to the first survey administration. Unique identifiers for each participant should also be written on the surveys for the participants.
- The tracking spreadsheet should be password protected and only one or two staff members should maintain and manage the tracking spreadsheet. These individuals should *not* be entering the data.
- Upholding ethical standards is of utmost importance when conducting research studies. Consent is a way to carry out these ethics.
- Passive consent is a type of consent commonly obtained in program evaluation; a participant can opt out of participating, otherwise, consent is assumed.
- DO NOT put the participants' names on the surveys or have the participants write their names on the surveys. Confidentiality must be maintained to obtain honest responses from participants.
- Proper conditions in the room during survey administration are also critical to obtaining accurate answers and setting a respectful tone.

## CHAPTER 6

### DATA RESULTS: “STATISTICS FOR DUMMIES”

---

Perhaps the most gratifying phase of the evaluation process occurs when you have the data results. If you are conducting a utilization-focused evaluation or partnering with an evaluator, you should ultimately have results, in a user-friendly format, to share with the staff of your program. Although it will be near-impossible to not think of your results as your agency’s “report card”, please keep in mind that any single evaluation, or any single measure, that shows undesirable results may be just as much (or more) about the evaluation as about the program itself. At the same time, you should not simply discount undesirable results as the fault of the chosen methodology. Positive results can provide a pitfall too, if we use them as a justification for complacency, or the thought that “our job is done”. Positive results should also be examined with a healthy skepticism to allow for the possibility that program success may be being overestimated. For example, if in a matched pre/post design only the highest functioning people in the program completed both pre- and post-tests, it will look like your program is more successful than it is. It is an exciting process to begin to “unpack” data results with both internal and external stakeholders, to inspire an examination of all possible avenues that will help you reach your ultimate goal of improving the quality and sustainability of your program.

Here are some definitions of common terms, especially those used in a matched pre/post design that will likely be included in your data results:

**Items:** the individual questions in a scale.

**Scale:** a collection of related items (questions) that measure a specific construct or attribute, for example, “Perception of Harm” is one scale on a survey about substance use and attitudes.

**Measure:** the entire set of questions – used interchangeably with instrument, survey, or tool. Sometimes used interchangeably with Scale.

**Sample:** the people included in your data set.

**Overall Mean:** the average score of all the items (questions) in the scale, across the whole sample.

**Item mean:** the mean score of one item (question) in a scale, across the whole sample.

**“R”:** when an item or survey question is marked with an “R” it means “reverse-scored”. To create overall means, some items may need to be reverse-scored, so that all items have the same expected direction of change. Usually, but not always, negatively-worded items are reversed so that in an overall mean, “more is better”.

**Pre:** reference to the data at baseline before program services.

**Post:** reference to the data after program services.

**Mean Change:** the difference between pre-test and post-test means

**P-value:** the probability that a difference between two means is due to chance alone. Smaller p-values are more desirable, indicating that the mean difference is large enough to call it a “real” difference.

**Program Effects:** what you would like the data to reflect as a result of your program (for example, an increase in perception of risk and harm of substance use.)

**Effect Size:** *how much* of an effect did the program have?

A paired samples t-test analysis examines the difference between the pre- and post-test means to determine whether an observed difference is statistically meaningful. The analysis calculates the probability (p-value) that the changes in the scores of the pre-test and post-test are due to chance rather than being a real change.

Standard practice in the social sciences is to consider p-values less than ( $<$ ) 0.05 statistically significant. This basically says that social scientists will only conclude that a difference between two means is meaningful when there is less than a 5% probability that the difference is due to chance alone. In some cases, especially when sample sizes are small, p-values between .05 and .10 are worth noting because they approach the .05 benchmark. In these cases, the term “approaching significance” is typically be used.

A limitation of interpreting the p-value and statistical significance is with regard to statistical power. Most small programs lack an adequate sample size to evoke confidence in the p-value and test of statistical significance. In the case of a sample size of 15 or fewer, p-values should not be interpreted. In those instances, effect sizes can provide a “benchmark” for comparison against other small sample results.

Effect size is another standard practice used in the social sciences. Calculation of an effect size can answer the question: *how much* of an effect did the program have? Effect size analyses provide an indication of the amount of change regardless of sample size. Effect size can be interpreted similarly to a “percent difference”, on a metric between .00 and .99. Effect sizes can be negative or positive, and a score of 0 represents no change. Generally speaking, effect sizes in social research are likely to be small (under .2).

The following is a set of benchmarks to determine the magnitude of effect sizes for your program:

.20	=	small effect size
.50	=	moderate effect size
.80	=	large effect size

It is possible for an effect size to be greater than one if the size of the effect is greater than one standard deviation from the mean. This sometimes happens when program measures are designed to fit exactly with a program curriculum, or when a measures assesses developmental change (i.e., growth), which happens rapidly in children whether or not they are participating in a program. Effect sizes and p-values can be used together to provide a more comprehensive picture of true program outcomes

### **What does it mean if my results are statistically significant (or if the mean differences have large effect sizes)?**

Statistical significance and/or moderate-to-large effect sizes might mean that your program is doing what you hoped it would. However, it is important to consider whether other factors, unrelated to your program, might have had an impact on the results. For example, if the change in bonding (as measured by the mentoring pre-test and post-test) was desirable, it is possible that your services were part of the reason for the change. However, it is also possible that these youth experienced something else during this time period that also had an impact on their bonding with school and community. For example, perhaps their school just received additional funding for new

facilities and teachers and this influx of financial support translated into more interesting and exciting programs at the school. While you want to celebrate the success, it is always important to interpret statistical results with caution. Most importantly, continually think about ways to improve your evaluation-related practices as well as program services.

### **What does it mean if my results are NOT statistically significant (or if my effect sizes were very small)?**

Remember, guidance on statistical significance applies only to those agencies that have sample sizes of 15 matched participants or more. If your agency has fewer than 15 matched participants, you are unlikely to demonstrate statistical significance *due to small sample size*.

Just as it is important to use caution when attributing statistical significance or large effect sizes to your program, it is important to understand that, sometimes, positive effects are difficult to detect. If some evaluation results are *not* statistically significant or have a small effect size, this does not necessarily mean that your program did not cause change in your participants.

Potential reasons may be:

#### Evaluation Implementation

- Your sample size (the number of participants who took your pre-test and post-test) was too small for a change to be detected. Did you accurately track participants from pre- to post-test? Did too many participants drop out (attrition) from your services? Did you lose data due to data-entry error?
- Participants weren't completely honest in their responses. Did you ensure confidentiality? Did you make sure participant names were not included on their surveys? (Be sure that you have a tracking system in place!)
- Your measure was not the best fit to capture the change. Is it reliable and valid? Does it measure outcomes directly tied to your services? Does it provide a range of response options?

#### Program and Population

- Your program may not be sufficient to make a significant effect on the participants. Do you want to consider adding or supplementing services? Do you have the right combination of services? Were you implementing the program with fidelity? That is, did you deliver enough of the program in the way it was intended?
- You did not target the population you intended. For example, you are measuring change in attitudes towards use. As it turns out, youth in your program already come in with very high negative attitudes towards use – meaning that there is no room for even more negative attitudes. Thus, there was no room, mathematically speaking, for statistically significant change from pre- to post-test. (This is often called a *ceiling effect*.) However, you thought you had targeted higher-risk youth.

Thus, result interpretation should be carried out within the context of the program. This means that you and your program staff should think about what happened over the course of

the program that could be an explanation for the data results, both positive and negative results. Process evaluation information (see Chapter 2) could help in this process.

### Chapter 6: Take-Home Points

- The interpretation of data results is an intricate process involving not only the numerical outcomes, but also program factors.
- Your evaluator should give you the data results in a user-friendly format that is easy to read and understand.
- All data interpretation should be done with caution.
- P-values should not be interpreted for sample sizes under 15 matched participants.
- Effect sizes and p-values can be taken into consideration together to provide a more comprehensive picture of true program outcomes.
- Other programmatic or contextual factors may be possible reasons that your results are not what you expected. Evaluators typically do not know and therefore cannot account for these factors unless they have been tracked carefully in a process evaluation.

## CONCLUSION

---

For program evaluation to be a smooth, ongoing process, all members of your agency need to be invested and recognize the benefits to the program and, ultimately, the clients being served by your agency. Remember that evaluation is *systematic*, meaning there are steps throughout the process that are thought out, planned, and then carried out in a methodical manner in accordance with accepted research principles. Thus, the earlier you can start thinking about and planning your evaluation efforts, the greater the likelihood of a smooth and efficient evaluation process.

The initial step is to have an understanding of evaluation components and concepts. Creating a logic model is the next step. The development of a logic model is a good method of conceptualizing your program and then moving to formulating and articulating more concrete goals, and proximal and distal outcomes. This may seem like the most challenging step in the evaluation process because the various stakeholders may have different ideas about the goals and outcomes of your program.

Once the logic model has been created, you can begin to plan your process and outcome evaluation endeavors. Remember that the process evaluation should always include some form of “dosage” information, that is, “how much” of the program each individual is receiving. It can also include data such as basic demographics on program participants and staff, client satisfaction surveys, or “live” observations of a curriculum as it is being taught. Measure selection to capture your outcome evaluation can be easy if you are required to use a certain measure as mandated by a stakeholder, or it can be a more demanding search. Remember to balance breadth and depth of the measure you chose to assess your program. Measures should not be expected to capture every outcome of your program. Often, focusing on one or two key outcomes can be helpful as well as convincing to stakeholders, especially if you can document links between those proximal outcomes, and other, longer-term outcomes.

Survey administration and participant confidentiality are intertwined in that you must protect participants’ confidentiality while administering the surveys. Delegating one or two staff members to manage a password-protected participant tracking spreadsheet is an easy method of upholding participant confidentiality. If people at your agency are responsible for entering the data, those individuals should be different from the ones who have access to the tracking spreadsheet. No one at an agency should ever match names with surveys once they are received.

Data results and interpretation are the final and most exciting steps. These steps assist in program improvement because they reveal areas of strength and potential growth within your program. However, be cautious when interpreting the data results. Contextual program factors can play a role and process evaluation information can be helpful to determine why you got the results, whether positive or negative.

Hopefully, this manual has helped you determine how to incorporate evaluation into your program in the most practical and efficient way, or has helped you think more deeply about elements of your evaluation already in place. Remember, the philosophy of, “How can we function well without knowing how well we’re functioning?” is essential, so that all parties understand why evaluation is useful and don’t feel that it is something unnecessary, simply being imposed upon them. Similarly, both program staff and evaluators are responsible for making the evaluation process and results useful, since we all have in mind the same goal of improving the social condition and therefore people’s lives.

## References

- Administration on Children, Youth and Families. (1994). The program manager's guide to evaluation (contract number 105-93-1915). Washington, DC: Department of Health and Human Services. URL pulled 1/1/2007:  
[http://hslcdev.cleverex.com/hslc/Program%20Design%20and%20Management/Program%20Design%20and%20Management/Communication/Performance%20Management/manager\\_pub\\_00021a1\\_072905.html](http://hslcdev.cleverex.com/hslc/Program%20Design%20and%20Management/Program%20Design%20and%20Management/Communication/Performance%20Management/manager_pub_00021a1_072905.html)
- Clegg & Associates, Inc. (2006). URL pulled 12/5/2006:  
<http://www.cleggassociates.com/Resources/LogicModel/index.asp>
- Israel, G. (2006). Using logic models for program development. University of Florida, Institute of Food & Agricultural Sciences (UF/IFAS). URL pulled 12/5/2006:  
<http://edis.ifas.ufl.edu/WC041>
- Khamsi, R. (2006). Close pregnancies linked to premature births. NewScientist.com.  
<http://www.newscientist.com/article/dn9015-close-pregnancies-linked-to-premature-births.html>
- McNamara, C. (1998). Basic guide to program evaluation. Free Management Library, Authenticity Consulting, LLC. URL pulled 12/5/2006:  
[http://www.managementhelp.org/np\\_progs/np\\_mod/org\\_frm.htm](http://www.managementhelp.org/np_progs/np_mod/org_frm.htm)
- McNamara, C. (2006). Guidelines and framework for designing basic logic model. Free Management Library, Authenticity Consulting, LLC. URL pulled 12/5/2006:  
[http://www.managementhelp.org/np\\_progs/np\\_mod/org\\_frm.htm](http://www.managementhelp.org/np_progs/np_mod/org_frm.htm)
- Rossi, P. H., & Freeman, H. E. (1993). Evaluation: A systematic approach (5th ed.). Newbury Park, CA: Sage Publications, Inc.
- Royce, D., Thyer, B.A., Padgett, D.K., & Logan, T.K. (2001). Evaluation design: an introduction (3<sup>rd</sup> ed.). Belmont, CA: Wadsworth/Thomson Learning.
- Schmitz, C.C. & Parsons, B.A. (2006). Everything you wanted to know about logic models but were afraid to ask. URL pulled 12/5/2006: <http://www.insites.org/documents/logmod.htm>
- The University of Wisconsin-Extension website. (2006). URL pulled 12/5/2006.  
<http://www.uwex.edu/ces/pdande/Evaluation/logicmodels.htm>
- University of Arizona Cooperative Extension, College of Agriculture & Life Sciences. (2006).  
<http://cals.arizona.edu/extension/pda/logicmodel/intro.html> (12/5/2006). Credit to: Allison Titcomb.

## **APPENDICES**

## APPENDIX A

### Evaluation FAQs

The following section contains questions often heard from program staff engaging in evaluation activities, and answers to those questions, as well as where in the manual more information can be found.

#### **1. Why can there be the same mean change from pre-test and post-test on two different items and one is statistically significant and one is not (or one has a larger effect size than the other)?**

*Answer:* The reason this typically happens is because of the different sizes of the *standard deviations* for each of those items. When all the survey responses cluster tightly around the mean (e.g., scores of 2.9, 3.0, and 3.1 result in a mean of 3), the standard deviation will be very small, which gives the equations “confidence” that the mean really is representative of the responses given by the participants. When the survey responses are spread across the whole range of response choices (e.g., scores of 1, 3, and 5 result in a mean of 3), the equations have to take into account that the resulting mean is not as representative of the scores. Thus, you find weaker results, even with the same exact means, when the standard deviations of those means are large.

#### **2. What steps are actually involved in compiling and analyzing the data?**

*Answer:* Once you collect surveys from your participants, the process of data entry and analysis will differ depending on your evaluation plan. In most cases, you will enter the data into a spreadsheet program, such as Access or Excel. The “data shell” into which you enter the data usually will be created by the evaluator or analyst. Then, a data analyst will later have to import the data into a statistical software package, such as SPSS or SAS. Once imported into a statistical package, the data have to be “cleaned”. Analysts will examine data for issues such as duplicate data or too much missing data, and may delete cases if they have reason to believe that the data are there in error (e.g., from the wrong year of the program). Data analysts do not have preconceived ways of assessing whether individual patterns of answers are bogus – that is what mathematical estimates of *error* are for, which the statistics take into account automatically. Contrary to fears that you may have, dishonesty or not taking surveys seriously do not typically present a problem in obtaining accurate results. Once the dataset is ready, analysts perform statistical analyses, such as paired samples t-tests. The results are outputted and the critical findings (e.g., means, standard deviations, significance levels) are extracted by the analyst and put into data tables that can be easily interpreted by a non-statistician.

#### **3. What if there isn’t an instrument that accurately measures my intended outcomes?**

*Answer:* The chances are that you either need additional assistance in locating a measure, or, you may need to redefine your outcomes so that they are more proximal and therefore possible to measure in the first place. See Chapter 4 for more examples of how outcomes relate to measures.

#### **4. What do I do if the participants do not understand the language in the questions?**

*Answer:* If you know that your program serves individuals who do not speak English, it is important to dedicate some of the evaluation budget to translation services. Ideally, paper measures should be translated, in written form, by certified translators. If this is not possible or available yet, a translator can work with a participant individually to verbally translate the questions and answer choices. If external stakeholders want an accurate picture of your program, the evaluation must include all participants, regardless of primary language spoken. This can be part of your overall argument to your funders about the resources you need to be able to provide a high quality evaluation.

If participants are English-speaking but still do not understand what a question is asking, it is not a good idea to offer your interpretation. A better idea is simply to say, “Do your best to understand the question. There is no right or wrong way to answer it”. Often, the reality is that the participant is hoping for guidance on how to answer the question – it’s not that they truly don’t understand it. If you feel that the survey is on a reading level that is beyond *most* of your participants, this is a larger issue that needs to be discussed with your evaluators. In rare cases, you may need assistance in finding a measure that is worded at a more appropriate reading level.

#### **5. Can I change a word or a question on the instrument?**

*Answer:* As discussed in the previous answer, it is not appropriate to offer your own wording for a question, either verbally or in writing. That is because small changes in wording often have a much larger impact on people’s answers than you might suspect. Measures have been validated on the specific word choices and the results cannot be said to be parallel if the words are different. If you have concerns about specific questions, discuss these with your evaluators, but not your participants. See Chapter 4 for a discussion about the reasons for negatively-phrased questions.

#### **6. How do I administer pre/post-tests if I serve people over multiple years/program cycles?**

*Answer:* The answer to this question will depend on the evaluation plan and how the data are analyzed. So long as participants are not being unreasonably burdened, it never hurts to “pre-test” them again in a subsequent year or program cycle. The critical question is how your analyst will analyze the surveys from individuals with more than one pre- and post-test. Most funders operate on a fiscal year basis and expect results to be demonstrated within one 9- or 12-month period. This is usually not a problem because most agencies do not have enough repeat participants to skew the results. Generally speaking, when analyses do include multiple years or program cycles, analysts will select the earliest pre-test and most recent post-test, for each individual. Analyses across more than two time-points (i.e., tracking progress over three or more surveys) are rare in program evaluation because they are only meaningful if *most* of the participants returned surveys at *all* time-points being examined.

#### **7. How do I administer pre/post-tests to comply with a fiscal year evaluation when my program started in one fiscal year and ended in another?**

*Answer:* Typically, participant data are included in the analyses for the year in which their *post*-test or *exit* survey was collected. If you have a program that is at least 5 months long, and it ends just after the fiscal year, it is ok to administer the post-test or exit survey prior to program end. You are the best person to make the judgment call as to which benefit is more important – having additional

participants in the current fiscal year, or making sure 100% of the program content is given before the post-test is administered.

## **8. Do we have to obtain consent forms from adults on themselves and their children?**

*Answer:* Program evaluation surveys are intended to provide information about programs - not generalizable, scientific statements about human beings. For this reason, ethical standards in research stipulate that evaluation surveys are an optional part of program activities and do not require ethics board-approved consent forms. Passive consent is usually acceptable – see Chapter 5 for a detailed discussion and Appendix B for a sample passive consent form.

Your agency may be under stricter rules for obtaining participant permission, particularly if you are a “covered entity” according to HIPAA laws. Typically, any agency that has access to health records is considered a covered entity. As a matter of course, the program directors at your agency should discuss whether and how HIPAA regulations may apply. The following website provides initial guidance about how to determine whether you are a covered entity:

[http://www.cms.hhs.gov/HIPAAGenInfo/06\\_AreYouaCoveredEntity.asp](http://www.cms.hhs.gov/HIPAAGenInfo/06_AreYouaCoveredEntity.asp)

## **9. Is there a “magic” number for a sample size to show significance?**

*Answer:* Statistical significance is greatly affected by sample size, but it is also affected by other things, such as standard deviation (see Question 1 for additional information). So while there is no magic number for sample size, we recommend that when your sample size is 15 or fewer, effect size, rather than statistical significance, should be used to interpret the finding. That is because very small samples do not have sufficient power to detect a significant difference, even if present. Conversely, very large samples will detect a very small difference as statistically significant. Just because the math has “confidence” that a difference is not due to chance does not mean we have found a difference that is truly meaningful in people’s lives – a difference known as *clinically significant*. Therefore, even with larger samples, effect size and statistical significance should be considered together for a fuller picture of the results. See Chapter 6 for more details about possible reasons for both significant and non-significant findings.

## **10. I have multiple funders with multiple evaluation requirements. Can I use the same measure to satisfy more than one funder? What do I do if every new amount of funding has another evaluation requirement associated with it?**

*Answer:* If you are reading this question and answer, it probably means that you are already familiar with the complexities involved. You should know that most agencies struggle with this issue. Streamlining of funding and evaluation goals is occurring, albeit slowly, on the policy level. In the meantime, the best thing you can do for your own agency is to “arm yourself” by having a high-quality evaluation already in place. The more your agency knows and can show about evaluating your own effectiveness – systematically and with the accepted methodologies that have been discussed throughout this manual – the more likely funders will be respectful of your process, and make exceptions to requirements, if necessary. Chances are you have a good argument for why it makes sense to limit the number of measure(s) you administer and to have the same measure(s) satisfy more than one funder. Consult with your evaluators to help you make this argument.

### **11. How do we know how our data compare to the state and national averages?**

*Answer:* The availability of aggregate data is one of the great advantages of taking part in a larger-scale evaluation with an approved list of measures. If this is the case, your evaluators should provide you with whatever comparison data are available for the same measures you use. If data from the exact same measures are not available, you and your evaluators can still do research to find out about what baseline data exist for similar constructs/outcomes at the local, state, and national levels. Local agencies often publish their annual reports online, which can give you a good idea of what other agencies like yours are expected to report to their stakeholders. Typing the name of your outcomes into a search engine will likely result in a number of sources with associated data reports.

### **12. What is the difference between outcome and process data?**

*Answer:* Simply put, process data is about your program, and outcome data is about the clients who take part in your program (see Chapter 2 for a detailed discussion and examples). The bottom line is that the right questions have to be asked, and the answers have to be tracked in the right way, in order to be able to make use of *either* type of data. You cannot provide answers about how staff education level (process data) impacts your effectiveness (outcome data) without systematically tracking staff education level, carefully defining your effectiveness outcomes, and logically linking which aspects you expect to be related to which other aspects. You are most qualified to think about the linkages that make your program work; your evaluators can help you systematize your questions and methodologies in ways that are likely to produce the most meaningful information.

### **13. I do not have an internal evaluator or training in research or statistics. Should I or someone at my agency take a course to be able to conduct analyses ourselves? Isn't the ultimate goal for agencies to take ownership of the evaluation process?**

*Answer:* In the same way that a data analyst could not simply take a course and be qualified to administer the services that you do, only people with years of advanced training in methodology and statistics are qualified to conduct analyses. Just as the consequences of having unqualified individuals administer social services could be potentially grave, the consequences of having unqualified individuals analyze and interpret data could be potentially grave too. While the mechanics of telling a computer to run a test may seem fairly straightforward, the specialized skill required is independent of the mechanics themselves, just as the intricacies of your interactions with clients and communities cannot be boiled down to step-by-step formulas.

Program administrators can have full ownership of the evaluation process without being data analysts in the same way that a person building a home can have full ownership of the process without being an engineer or architect. Of course, a certain amount of mutual trust is necessary. It should be obvious whether there exists mutual trust and respect for each other's professional domains. You *should* monitor your analysts, whether internal or external, by asking questions and always holding them to the standard of making their products meaningful to you.

**APPENDIX B**

**Sample Informed Consent Form**

*THIS SAMPLE, PASSIVE CONSENT FORM IS PROVIDED FOR YOUR CONVENIENCE. CHECK WITH THE APPROPRIATE PERSON(S) TO SEE IF PARENTAL OR ACTIVE CONSENT IS REQUIRED.*

We would like you to participate in the evaluation of *[program name]*. Your participation is important to us and will help us assess the effectiveness of the program. As a respondent in *[program name]* we will ask you to *[complete a questionnaire, answer questions in an interview, or other task]*.

We will keep all of your answers confidential. Your name will never be included in any reports and none of your answers will be linked to you in any way. The information that you provide will be combined with information from everyone else participating in the study.

You do not have to participate in the evaluation. Even if you agree to participate now, you may stop participating at any time or refuse to answer any questions. Refusing to be part of the evaluation will not affect your participation or the services you receive in *[program name]*. You only need to hand in this form if you want to DECLINE participating in the evaluation surveys. Otherwise, it is assumed you have no objection to filling out the surveys.

If you have any questions about the study you may call *[name and telephone number of evaluator, program manager or program staff member]*.

DO NOT AGREE TO PARTICIPATE

**Signed:** \_\_\_\_\_

***Respondent or Parent/Guardian***

**Date:** \_\_\_\_\_

## APPENDIX C

### SAMPLE SURVEY INSTRUCTIONS

*To be Read by the Survey Administrator*

**[PLEASE MAKE SURE YOU HAVE RECORDED EACH RESPONDENT'S NAME ALONG WITH THE UNIQUE ID ON A TRACKING SHEET PRIOR TO SURVEY ADMINISTRATION.]**

**[PASS OUT A SURVEY TO EACH RESPONDENT. ASK RESPONDENTS TO PAY ATTENTION TO THESE INSTRUCTIONS AND NOT TO BEGIN THE SURVEY UNTIL YOU HAVE ASKED THEM TO DO SO.]**

**[PLEASE READ TO THE GROUP:]**

Today, we will be completing the [survey name] Survey. **DO NOT** write your name anywhere on the survey. Your survey is identified with a number so as to protect your privacy.

This is not a test and there are no right or wrong answers. It's important to understand that your answers are completely confidential-- that is, no one at [name of organization] nor anyone else will ever know your individual answers. We are asking for your time today in an effort to improve the services we offer, which will require looking at everyone's answers combined.

Please answer every question as honestly as you can so the information that comes from this survey is correct and useful. Try to complete all items in the survey as best as you can, but if you come to a question in the survey that you do not want to answer, you can leave it blank and move to the next question. If you don't find an answer that fits exactly, use the one that comes closest. All of the questions should be answered by filling in one of the answer spaces. If you have a question about an item, please raise your hand. If you feel that a question does not apply to you, leave it blank.

Your answers are very important and only YOU can describe your experiences and opinions rather than rely on others to describe these things for you. You may choose not to complete the survey.

*[Are there any questions?]*

*[Please begin completing the survey now.]*