HIV Mental Health for the 21st Century

Winiarski, Mark G.

Published by NYU Press

Winiarski, Mark G.
HIV Mental Health for the 21st Century.
Project MUSE. muse.jhu.edu/book/15730.

For additional information about this book
https://muse.jhu.edu/book/15730

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=499525
In these times of shrinking resources and cost containment, funders and providers of HIV-related services want more than ever to know that they get results for dollars spent. In fact, the pressures to demonstrate whether health service programs have a measurable effect are probably greater than at any other time in recent memory. Until the political pendulum swings back to the days of "The Great Society," budget cuts will most likely affect those service areas that fail to demonstrate that they make a difference.

Government funding agencies and foundations now require that sophisticated evaluations be conducted to determine whether mental health and other program dollars are well spent and whether new models of care being tested or "piloted" through grants should receive additional attention.

Ethical providers of HIV-related services ask similar questions: "I know that people like our mental health clinic. But does our psychotherapy, or our psychiatric service, really improve people's lives?" and "Are there ways of finding out how we can improve our services?"

Program administrators who seek quality and patient satisfaction rely more and more on evaluations. For them, evaluations feed back information to service providers so that they can improve their programs. Evalua-
tions also help define which services work better than others and provide guidelines on how to make service packages most effective. For programs, a good evaluation creates an information loop, informing administrators and clinicians, who then modify the service package, which is again evaluated.

Not too long ago, many evaluations were based simply on surveys of either a program’s providers or its clients at one or several points in time, asking them questions about the program’s effects. This strategy is weak for many reasons. It is based on subjective notions about what works or doesn’t work, and it fails to provide concrete evidence for its conclusions. It does not consider how things were before the new program began, or even at different stages of the program, and it does not consider that respondents may have many reasons to praise a program, only some of which may have to do with real effects. Employees may say wonderful things about a program, for example, so that they can keep their jobs, or clients may praise a program so that they don’t risk angering their service providers.

These types of evaluations are no longer considered sufficient. The field has become much more sophisticated: A profession of evaluation now exists, and professional societies have sprung up. Many consulting firms offer evaluation services, and some academic departments have evaluation experts on staff.

This chapter, and the next, provide basic information about evaluation, with an emphasis on HIV-related issues, so that readers can enter the mental health field knowing the questions it is asking itself and how it is attempting to find the answers. If any reader ever becomes involved in requesting money from governments or foundations, this information will be a start toward understanding what must be included in a viable application for grant funds.

This chapter focuses on quantitative evaluations, that is, those done with numbers and statistics; I illustrate my points with an actual federally funded HIV mental health project. The next chapter discusses qualitative evaluations.

What Is Evaluation?

In brief, an evaluation seeks to answer questions about a program’s processes and outcomes. Typical process questions, broadly stated, are: Are we reaching the service population(s) we hope to serve? Is the full
array of psychosocial problems being addressed? Qualitative evaluations are more likely to deal with process, whereas quantitative evaluations typically deal with outcome questions, including these concerns: "Does the program make a difference?" and "Is the program cost-effective?"

Many evaluations are concerned with outcomes that can affect local or national health care policies, since many governmental and foundation grantors want applicants' projects to be generalizable and replicable. Generalizability means that findings determined at one agency can be applied at— are generalizable to— others. Federal agencies often fund tests of new programs not because they want to assist one agency but because the findings are potentially useful to hundreds of agencies. Replicability means that the structure of the program being studied, if successful, can be adopted at other institutions and will provide the same results. Programs that are very idiosyncratic, and hence not replicable elsewhere, are not likely to receive funding.

Going from these broadly stated questions and concepts to meaningful specific questions and answers is a very rigorous process that requires evaluation expertise and involves labor, time, and a willingness to tolerate the possibilities of non flattering results. If you learn nothing else, learn that you need to seek the assistance of experts before you plan the program. The program and evaluation should be built together.

CASE EXAMPLE: Congress, in passing the Ryan White Comprehensive AIDS Resources Emergency (C.A.R.E.) Act, created the Special Projects of National Significance (SPNS, pronounced "spins") program to support development and evaluation of cutting-edge programs in HIV services. SPNS programs were created at Montefiore Medical Center in the Bronx, New York, and at St. Joseph's Hospital and Medical Center, in Paterson, New Jersey, to integrate HIV-related mental health services with primary medical care. My task, and that of my colleagues in the research division of the Department of Family Medicine, Albert Einstein College of Medicine, was to evaluate those HIV-related mental health programs.

How to Think about Quantitative Evaluation

A quantitative evaluation can best be described as a comparison or, more accurately, as a whole host of comparisons.

The most common comparisons to consider for a new program are:

- The old way of doing things versus the new program.
- The new program versus a "control" program. A control pro-
gram is one that appears very much like the new program but is absent the "active ingredients" — those that are believed to make the difference — contained in the new program.

For a quantitative evaluation of a mental health program, patients are generally assessed in some way both before and after being exposed to the program. This comparison is called pretest versus posttest. The program manager's hope is, of course, that the programs will be found to have made a difference in the client's life.

Another comparison that can be made is one made across time. The scores of tests taken in January, April, July, and October, for example, may be compared to see if there is a pattern of change. This is called a time series and is really just an extension of the pretest-posttest idea but with multiple measurement points.

Those Confounding "Confounds"

If these comparisons were a simple matter, experts (and this chapter) would be unnecessary. Unfortunately, making comparisons that stand up to scrutiny can be devilishly tricky.

Confounds (some people call them confounders) are factors that undermine the comparisons made in the evaluation process. Here are some confounds (there are many others):

- Regression to the mean. This concept refers to the fact that individuals usually enter treatment at personal low points (e.g., when severely depressed or anxious). Since these emotional states are variable within individuals, a reassessment later will, on average, simply measure a natural rebounding of the mental state. This rebound would be expected to occur naturally whether or not the person received treatment. This is referred to as "regression to the mean" because it is a measure of a return to an average (mean) mental state.

- Other changes that occur during the course of the program that have an effect, independent of the treatment itself, on the desired outcomes. The person being treated for depression, for example, may lose a partner or be evicted from an apartment. His or her depression scores may then worsen, despite the clinician's best treatment efforts.

- Participation in multiple programs. A client may attend several
different treatment programs in the community so that it cannot be ascertained that one program alone was responsible for improvements.

- Changes in treatment agent. A patient's therapist may leave the program, for example, negatively affecting the client.

In evaluation, these are referred to as “historic” or “contextual” biases. Campbell and Stanley (1963) list many other confounds, which they term threats to a research design's internal or external validity.

A successful evaluation requires that as many of the confounds as possible be eliminated through the study design or controlled for in the analyses by an examination of their independent effects. If serious confounds cannot be eliminated, it may not be possible to learn whether a program makes a difference, and funders may respond negatively to an application for funds.

**What Do I Evaluate?**

One of the most challenging aspects of creating an evaluation plan is deciding what aspects of a program should be assessed. The best way for you as the researcher to think about this is to reflect on these considerations:

- What groups do you want to affect? "Clients" may be your immediate answer. But do you want the program also to change the practice of clinicians? Do you want the program to change the institution or agency in some way? If so, then you need to evaluate whether these changes occurred.

- For each group, what change do you want to occur as a result of and only as a result of your program? For clients, you may want to determine whether your services affected their depression. For a clinic, you may want to increase physicians' referrals to mental health specialists.

**Case Example:** In the Montefiore and the St. Joseph's programs, providers felt that patients receiving mental health services would generally feel better, that they would have better coping skills, that they might not show up in the emergency room as often, and that their feelings of stress would be reduced. It was also felt that in those clinics with integrated mental health care, a greater number of patients with HIV would avail themselves
of those services since they would be available on site. This arrangement would surmount the barrier of traditional reliance on referral to another agency or provider at another location. The program developers also believed that medical providers might feel better about their work if they could make quick mental health referrals and receive quick consultations.

**How Do I Measure These Things?**

Once you decide what aspects you want to measure as part of the evaluation, you will want to consider how to measure the attributes. Some would say you have to “operationalize” those aspects—to give them concrete, measurable meaning. That concreteness generally is stated in the phrase “as measured by [name of instrument].” Instruments use numbers to express the concept being measured. These numbers are likely to include raw scores (the client’s score without any statistical massage); corrected scores, which are raw scores massaged, for example, to correct for response omissions; and scale scores, which are a measure of how the patient performed relative to a group. We call what we do quantitative evaluation simply because we use numbers—scores—as the bases for our comparisons.

The first task is to conduct a literature search of how others have measured the phenomena you want to measure. Many published scales exist that may suit your purposes, but the literature may indicate a preference for the use of one particular instrument. You may also want to consult the latest edition of what is familiarly called *Buros’ Mental Measurements Yearbook* but is officially *The 11th Mental Measurements Yearbook* (Kramer & Conoley, 1992), which reprint reviews of psychological instruments.

You will want to find an instrument that has proven *validity* and *reliability*. These words have special meaning to evaluators:

- A scale is *valid* when it has been proven to measure what it claims to measure.
- The scale is *reliable* when it has been shown to produce results that are the same when confronted with the same amount of the phenomenon. An instrument scale is reliable when the same results are obtained from the same individuals, provided they have not changed, in repeat administration of the instrument. Your bathroom scale is reliable if it is always ten pounds off, because the amount of error is always the same.
To ensure reliability, often those who administer the assessment instruments have to be specially trained to a certain standard, which is statistically checked, and they are rechecked over the course of the project. If you decide to use instruments that are administered by someone or several someones, such as psychologists, you will have to spend the time and money to train these individuals to a certain standard and to check them occasionally to ensure that the reliability is maintained.

Another important term is construct validity, which refers to the degree to which question items that measure different aspects of a construct "hang together" in a meaningful way. If, for example, you wish to measure a particular attitude, it is often important to have several questionnaire items that are intended to measure different aspects of the same attitude. The degree to which individual responses to the varied items follow the same response pattern is a reflection of the validity of the scale.

It is beyond the scope of this chapter to undertake a thorough discussion of the ways that the validity of an instrument is assessed, but a psychometrician should be consulted when planning an evaluation that uses assessment instruments. See also Campbell and Stanley (1963) for a discussion of different concepts of validity.

A major concern of those who work in the inner city is whether instruments are available that have been validated with minority culture groups, including groups that cannot read English. Unfortunately, even major publishers of assessment instruments have merely translated instruments into, for example, Spanish, without conducting new validation studies to determine whether the instrument really measures the desired phenomenon in the responding group. Issues of validity for personality assessment and psychopathology are well reviewed by Rogler, Malgady, and Rodriguez (1989) and by Marin and VanOss Marin (1991). Rogler, Malgady, and Rodriguez (1989) suggest that we should assume that there is cultural bias in instruments, unless proven otherwise. Certainly, an evaluator should be able to defend his or her choice of instruments as consistent with the American Psychological Association's Standards for Educational and Psychological Tests (1985), which offers many warnings to consider in work with cultural minorities.

In the event that the literature search does not produce an instrument with adequate reliability or validity, you might wish to produce a scale tailored to the particular issues of your evaluation. This can be a laborious process, but it can be quite enlightening as well.
One approach might be to conduct a focus group, in which a group of professionals gathers to discuss the meaning of a concept and to suggest the components of how it should be evaluated. The members of the group might suggest items to be used in instrument development. The rationale for having several participants is to make sure that many perspectives are included and that a sharper definition of the concept will result. It may even be worthwhile to have a focus group with the recipients of care to discuss how and in what ways they value the services. This may identify issues that weren’t thought about by the providers. A more complete discussion of this process of evaluation can be found in Chapter 18.

CASE EXAMPLE: In the evaluation of the integrated mental health services at Montefiore and at St. Joseph’s, several informal focus group sessions took place to discuss the qualities that should be included in the evaluation plan. We decided that patients’ sense of well-being, levels of depression and anxiety, dependence on emergency room visits as a result of crises in their lives, and levels of physical and social functioning might all be affected in a positive sense if they received mental health counseling in a timely and appropriate manner. A literature review was conducted to ascertain what instruments were already available to assess these issues. We uncovered two instruments that dealt with many of the psychosocial and physical functioning dimensions. These were the Medical Outcomes Survey (MOS) short-form questionnaire (SF-36) (McHorney, Ware, Lu, & Sherbourne, 1994; Ware & Sherbourne, 1992) and the Brief Symptom Inventory (BSI) (Derogatis, 1992; Derogatis & Melisaratos, 1983). Both scales have been used extensively and have been extensively assessed for their reliability and validity.

Another area we felt was important to include was the degree to which barriers to obtaining care would be altered. For this we found an instrument titled “Primary Care Physicians and AIDS Scale” (Gerbert, Maguire, Bleecker, Coates, & McPhee, 1991), which addresses this issue. Last, we thought that if mental health services were readily available to the health team, the program staff itself would feel a reduction in job stress. For this we located the Work Environment Scale (Abraham & Foley, 1984), which includes items that address how people feel about their jobs and the level of stress they perceive. All of these scales were available and finding them obviated the need to reinvent them. Figures for the other issues we felt to be important—severity of illness, use of the emergency room, the number of medical and mental health visits, etc.—were all obtained by reviewing patients’ medical records.
How Do I Structure the Evaluation?

One of the most vexing issues in doing a credible evaluation is deciding how to structure the evaluation or how to pick the most appropriate design for an accurate comparison. The best structure is one that eliminates all confounds and allows for an unambiguous response to the question “Did our program, and our program alone, make the difference?”

Scientific research methods have informed quantitative evaluations on ways to eliminate, as much as possible, confounds. The main strategies are these:

- **When making comparisons, compare your program with a control program, that is, a concurrently studied control group.**
  
  It may seem unethical to have a control group of severely depressed or anxious clients from whom you withhold treatment in order to obtain a “cleaner” measure of the effects of treatment. If so, then you are thinking good ethics. Whenever we deal with a clinical condition for which there is a known treatment that has evidence of a therapeutic benefit, it is unethical to have an untreated control group. We have to create a comparison of one treatment approach (the accepted standard) to another, the experimental approach. In this way, the regression-to-the-mean problem is handled because it presumably operates in both groups to the same degree.

- **When possible, randomly assign patients to your program and to a control program.**
  
  A design based on a random assignment to two or more groups is the ideal. This is the classic randomized clinical trial. In this design, patients in need of mental health services are randomly assigned to an experimental approach to treatment or to a traditional approach. Although this represents the standard one should always strive to approach, it is rarely done because of logistical issues. This approach has the best chance of overcoming all the problems or confounds inherent in other approaches because it assumes that confounding aspects are equally spread over each group. The nature and the severity of the health problems are similar in the two treatment groups, the historical factors that might have an impact on the outcomes are similar, and the
regression-to-the-mean issues are theoretically equal. In other words, by the randomization process, one has the greatest likelihood of creating two comparable groups in order to isolate the impact of one treatment approach relative to another.

**CASE EXAMPLE:** In the Montefiore evaluation, the goal was to assess the relative effects of a new delivery system of mental health services to HIV-infected individuals in the context of primary medical care. This new approach is called “integrated delivery of mental health services” because the mental health worker became a regular member of the health care team and participated on an equal footing with the other members of the team. The need for a mental health referral to another agency at another time was therefore eliminated. This modality of delivery of care was to be evaluated relative to the traditional model, which was based on referrals to other agencies or departments, all of which were at other locations. There were sufficient funds to develop an integrated service model in four of the eight Montefiore Ambulatory Care Network (MACN) sites. The remaining four, which would not be exposed to the integrated model, would be available to serve as control sites. In this manner, a “natural” experiment was possible in which an innovative approach to delivery of mental health services could be measured against a traditional model of care. This was not a randomized clinical trial because patients were not randomly assigned to one system or another, nor was the selection of sites for the integrated model made on a random basis. The best name for this design is probably a “cohort” model, because two cohorts were to be followed: the patients served by one of the two systems, the integrated mental health service model or the traditional delivery system.

We became aware of a potential problem relating to the regression-to-the-mean phenomenon. In the integrated model, patients would presumably more easily gain access to mental health services on average than they would in the traditional model based on physician referral. This might mean that at the outset of receiving services, those in the traditional model might be in worse states than those in the integrated model. In the reassessment of these patients at some later date, it would therefore appear that the patients in the traditional model had improved more, if only because they started at a different, lower point. To overcome this potential problem, we adopted an approach called an “intent-to-treat” model in which we evaluated mental and social functioning of HIV patients who entered both systems of care irrespective of whether they received mental health services. In this way we were able to address the issues of whether patients entering one system of care fared better relative to patients entering another system of care. Furthermore, we
were able to assess the degree to which mental health services were
delivered to more patients and whether the severity of patients’ mental
health problems was different at the point the patients received services.
The effects of care were assessed by comparing, in the broadest sense,
whether patients who entered one system of care fared better at some
later point than did patients entering another system of care.

**Compiling the Numbers**

You realize now that you will be doing a lot of measuring and that many
numbers will result. In addition, you will compile a host of information
on the clients themselves, including age, race, ethnic identification, sex,
sexual preference, route of infection, and other demographic information
you deem important. This descriptive information is important because
the reader of your evaluation report will want to know if your findings
will be applicable to his or her service population.

All these numbers have to be stored somewhere. This requires two
components: a database software package and a computer with enough
capacity to run the program. It is important also that you realize that
compiling and maintaining a database are very labor-intensive and require
a person willing to give a lot of attention to detail. Throughout the
project, the database manager has to ensure that the information entered
into the database is accurate and complete. Failure to attend to this
jeopardizes the entire evaluation.

**How Do I Analyze My Findings?**

This is not a chapter on statistics, which are uniquely complex in and of
themselves. To satisfy statistical requirements, for example, you must have
a sufficient number of clients in a quantitative analysis to be able to find
meaningful differences, if they exist. You must also determine what you
mean, statistically, by “difference.” And you must determine what param-
eters of error should be built into statistical equations. These and other
issues require the assistance of someone knowledgeable in statistics.

For purposes of introducing you to quantitative analyses, we intro-
duced the concepts of comparison and measurement. It is in the analyses
that these concepts interact and produce results. While statistics can do
many tasks, including making predictions, we will discuss only compari-
sions.
Comparing Group Means

We discussed evaluative comparisons, broadly speaking, and then the narrowing of these issues into operationalized, measured concepts. We now have a computer disk full of numbers, arranged correctly in designated rows and columns. We turn this disk over to a statistician who uses software, such as Statistical Package for the Social Sciences (SPSS) and Statistical Analysis System (SAS), which will compile the numbers into meaningful chunks. The statistician will have decided, on the basis of your initial proposal, what family of statistical tests he or she will use. Nonparametric statistics are used when certain kinds of data are used (e.g., frequencies), for example, and parametric tests are used when other types of data are examined (e.g., group means).

For evaluation purposes, perhaps the most typical comparison is that of group averages or means. The statistical program will compute the mean depression score of the group of patients enrolled in your new program and the group mean of all patients in the traditional program. The statistical software will also compute a measure of "score spread" for each of the groups. Then, the computer will take other orders—such as what you determined was a statistically significant difference and the error tolerance you will accept. Using these factors and the measurement data, the software will tell you if there is a significant difference between the group mean of your treatment patients and the group mean of your controls. This is the simplest comparison.

More complex comparisons are possible. Group means for many groups, or for multiple measurements of individuals in each group, can be compared using statistical processes such as analysis of variance (ANOVA). If you wanted to take many different scores—depression, anxiety, and sense of well-being, for example—and make multiple comparisons, the statistician might decide to use multivariate (multiple variable) statistics. A printout of a multivariate analysis might indicate that there were differences between the group means on some variables and not on others or on all variables or that no differences were evident, which would mean that your program functioned just like the program you compared it with.

Multivariate analyses are the minimum statistical analyses done in evaluation, because studies have to concern themselves with many variables that aren't even the target of our programs. These variables include gender, age, race, ethnic identification, and marital status, among many
How Do We Know It Works? | 287

others. If you want to learn if individuals in a certain age group or sex responded better to your program than did comparable individuals in the control program, then multivariate statistics are what you will use.

**CASE EXAMPLE:** The Montefiore and the St. Joseph's data, as compiled by the database software, are placed on a disk. Then statistical experts program SAS to respond with different analyses. One simple group comparison is whether patients at St. Joseph's Hospital use the emergency room fewer times than subjects in the control program, at another Paterson institution. A more complex analysis would be to compare use of the emergency room by Latinos or blacks in both study groups.

**HIV-Specific Issues in Evaluation**

When planning an evaluation of an HIV-related program, you need to take into account the following special concerns:

- Many inner-city clients are overwhelmed with myriad psychosocial and medical issues, and they tend not to put an agency's evaluation project on the top of their priority lists. Because they are unlikely to remain in a study that extends over several months, ensuring their continued involvement tends to be labor-intensive and expensive. Fowler et al. (1992), reporting on a longitudinal survey of persons with AIDS, noted that every facet of their work posed challenges; they suggested solutions to problems in identification and recruitment of subjects, consent, and data collection.

  More and more, evaluators and researchers are offering financial incentives to persons to induce them to complete assessment forms.

  **CASE EXAMPLE:** The Montefiore project research assistants spend significant time attempting to track clients so that they can complete additional assessment forms. Despite our efforts, there is a significant attrition between the first assessment and subsequent assessments. This requires that we enroll many extra patients, to compensate for the dropout factor. Also, we provided patients $10 apiece for each of three assessment sessions. They indicated that the money was a major factor in their participation.

- In evaluating services, you will likely need to include in your list of independent variables each patient's stage of illness. When
considering issues of depression and anxiety, for example, it may be that persons with more symptoms or in later stages of the illness have different levels of anxiety (perhaps more, perhaps less) than persons who are asymptomatic.

CASE EXAMPLE: We used the Centers for Disease Control (1987) staging system for HIV, which has four stages, with five substages in Stage IV.

Common Barriers to Evaluation

Before an evaluation can be considered, it has to be valued by administration and the service providers. If it is seen as an “extra”—something of limited importance—then the likelihood of its being treated as a serious activity or having a real impact on the program is doubtful. The mechanisms of how the results of the evaluation are to be used should be spelled out early in the planning process.

The value of early planning cannot be underestimated. Very often an evaluation is called for well after the program has started and when it is too late to measure things that were going on at baseline. This tends to render good evaluation designs less feasible.

I have found that behavioral scientists (e.g., psychologists, social workers) tend to resist using standardized instruments because, they feel, they limit their autonomy and expressiveness in conducting their clinical assessments. They also resent the “overly structured” feeling they get when they are asked to fill out standardized assessments. The problem the researcher faces is that different clinicians emphasize different aspects of a case, thus making it difficult to use narrative notes as a means to develop uniform measures of how patients are doing and what degree of change has taken place. One way to handle this common situation is to attempt to reach a compromise, often by conducting a focus group session with the clinical staff to identify the critical areas that all service providers should be addressing, and then adopting instruments for those areas and relying on narrative comments for the remainder of the clinical assessment.

If you don’t have the expertise to conduct a carefully designed evaluation, it may be worth contacting a local university or medical center to obtain help. For relatively affordable consultation fees, you may get the guidance you require, if not the assistance you need to develop a grant to
support a major evaluation effort. The academic community often welcomes the development of partnerships with service programs because such partnerships bring them into the “real world” to see whether their theories hold up.

Last, the evaluation process usually requires extra funding. Services are often covered by third-party payers, and the costs of the evaluation are not usually part of that package. Grant support, increasingly more difficult to obtain, is usually required. Many federal or state programs that fund clinical services, however, increasingly require evaluation activities to demonstrate the value of the services. In this case, it is appropriate to include the costs of evaluation in the program grant application. Even in cases in which the program announcement doesn’t specifically call for an evaluation, it is worth building in an evaluation component. It serves to remind the agency that you value evaluation and that competing renewals are more easily defended when there is good evaluation evidence of the program’s impact.

REFERENCES

This chapter was made possible by grant number BRH 970165-02-0 from the Health Resources and Services Administration. Its contents are solely the responsibility of the author and do not necessarily represent the official views of HRSA.


