



Chapter 3: Design and Methods

This chapter addresses a range of design issues you should consider in planning your performance measurement project. Although all of these issues have technical elements, many stakeholders, even those without technical backgrounds, will typically want to participate in resolving most of them. People with technical backgrounds involved in implementing the project should be expected to explain the underlying issues and practical implications of technical decisions to non-technical stakeholders so that the latter can contribute to making informed choices. These issues should be addressed in the planning meetings with stakeholders described in Chapter 2.

For each consideration, we will briefly elaborate on the *nature of and reason for* the issue. Next, we will describe how this issue appears in the *evaluation and research literature*. Then we will present the *major options* for addressing the issue. Finally, we will discuss the major *advantages and disadvantages* associated with the options.

As you will see, the advantages and disadvantages of options will revolve primarily around such issues as the *precision and consistency* gained or lost with an option, the *resources* required, and last but not least, the *burden and risks* to privacy for data providers. A general principle to keep in mind in considering options is that in performance measurement projects, methodological and design decisions will affect your ability to analyze data, interpret data, and act based on these data.

Along with the discussion of each issue, we have included references to other materials for those who wish to explore the topic in more depth. In some cases, we have included copies of particularly relevant materials in the appendices.

We organize our discussion of design and methods issues by the type of data to be collected. As discussed above, the indicators of the MHSIP Report Card are based on several types of data: the Consumer Survey and other self-report data, clinician assessments, enrollment and encounter data, medical records, and other administrative data. For the purposes of the discussion of design and methods, we will group these types of data into consumer reported (including the Consumer Survey and other self-report data), clinician ratings, and all other administrative data. But before discussing the issues regarding each data type, we discuss issues you will need to consider regarding content of your performance measurement system.

I. Content Decisions

This issue concerns the number of different aspects of system performance that will be covered in a Report Card project. At the most general level, these topic areas are sometimes referred to as *domains*. This term is used to describe the major sections of the MHSIP Report Card. In discussions of evaluation methods, this topic is often called *instrument content*.

The different domains covered by the MHSIP Report Card are described in Chapter 1. Each domain includes several indicators, each focusing on one particular feature of performance in the topic area. Depending upon the focus of a given report card project, not all domains must necessarily be measured, nor all indicators within a domain. However, if you want to be able to compare your data with MHSIP Report Card efforts implemented elsewhere, you should consider using as many MHSIP indicators as your constraints permit. The value of MHSIP Report Card data will be very much enhanced to the degree that efforts in different states have comparable datasets.

One set of options for this question can be stated generally in terms of more or less content. This is based partly on the level of resources that are perceived as being available for the project. Specific resources required for a Report Card project are discussed in Chapter 2. It is important to recognize that the decision about level of resource commitment to a performance measurement project is ultimately a political one, based, in part, on the perceived benefits of the resulting data.

Another important set of options for this question concerns the relative value and meaning of information about various dimensions of performance. In any given project, you must decide which domains (e.g., access, outcomes), and which dimensions of performance within any domain (e.g., cultural access, geographic access, functional outcomes, personhood outcomes), to emphasize. Different stakeholders can place different values on each of these dimensions. The effort required for data collection varies across both performance domains and settings, as does the value of evaluating performance in any domain, so each report card project must determine its own optimal balance of effort and focus.

One advantage of **more content** is that you have the opportunity to learn more about system services and outcomes, or to learn with greater precision within each domain. Another advantage of more content, assuming information is gathered on a wider range of topics, is that there is less chance of creating perverse incentives or allow for system "gaming." What this concept refers to is that, if financial or other rewards are based on narrowly defined performance measures, the actions that organizations take to show improved performance on those measures may have unintended, negative consequences or, at least, may be unrelated to overall system performance. For example, if the only performance measure for a system was the rate of re-hospitalization, it might create an incentive to hold persons who were judged to have any risk of readmission in the hospital indefinitely; thus, a system's score on this performance measure might or might not reflect effective hospital services, as it was intended to. A larger, more balanced set of indicators is less likely to distort system performance. On a more practical level, within a given project there is potential for some economy of scale. Once you have undertaken to ask a few questions or collect data in some other way, collecting a little more may require relatively little increase in resource consumption.

The disadvantage of more content is that it requires more resources at all levels of the system. Notwithstanding potential economies of scale, some expansions of content are more costly than are others. For example, expanding data collection by adding a completely new survey, or tapping an additional database, may require disproportionate resource expenditure. If you are using surveys, asking for more information can be more burdensome and intrusive for respondents. If the various components of the system do not have adequate resources to provide the scope of content desired, the data quality may be compromised, and the attempt to collect a wider range of information could actually have the perverse effect of leaving you *less* well informed. Proposals for outcome system standards have included recommendations to limit data collection to the scope that can be accomplished adequately within the resources available (Center for Mental Health Services, 1998).

As you consider which aspects of performance to examine in your project, it is worth recalling that the group of drafters of the MHSIP Report Card included representatives from multiple perspectives, and the domains and indicators recommended in the report represents a balance of values from these many points of view. If you think most broadly about performance issues in a large-scale behavioral health care system, and if you assume responsibility for service to diverse populations, you will typically choose to include multiple indicators within each of the domains of Access, Appropriateness, and Outcome at a minimum. Local conditions, including issues of both organizational evolution and political climate, will affect your choice of relative emphasis. Recent introduction of managed care structures may lead you to put special emphasis on critical aspects of Access. Particular concern for the effectiveness of programs serving adults with severe mental illness may lead you to emphasize recovery-oriented aspects of Appropriateness and Outcome. A general concern for interagency effectiveness in serving children may lead you to focus on MHSIP Report Card indicators relevant to this age group within a broader monitoring effort.

At this point of considering what aspects of performance to measure, it is important to consider both the technical and the practical or political value of including more than one perspective in your measurement design. Several early Report Card projects appear to have found it most feasible to implement only the Consumer Survey, deferring use of other methods. In many cases this decision may reflect the desire to give stronger recognition to a crucial and neglected perspective; in others the complexity of standardizing database collection has appeared more daunting than the task of implementing a large-scale but straightforward survey. But because interpretation of survey findings in isolation almost invariably leaves much

ambiguity unresolved in the minds of important audiences, including the sponsoring agency, some effort to include non-survey indicators at the outset of report card projects is recommended. The relevance of indicators drawn from sources other than the Consumer Survey is discussed in greater detail below as is the issue of whose perspectives to include.

II. Consumer Reported Data

As noted above, the major consumer reported data elements in the Report Card are the 40-item Consumer Survey, the Symptom Distress Scale, the Rosenberg Self-Esteem Scale, and a variety of other self-report items. The major design and methods considerations for collecting these data are sampling, reference times, timing and frequency for data collection, selecting an instrument, selecting a method of survey administration, and translation.

Sampling

Surveys are an economical means of determining characteristics of a population by observing a small sample of the population. Depending on the method of survey administration, it is possible but unlikely that you will have the resources to survey the entire population of interest; in most cases you will need to select one of several sampling strategies. The most basic form of a survey is known as a **simple random sample**. The value of this form of sampling is that, in general, the observed sample will have similar characteristics of the total population. More complicated forms of surveys include stratified random samples and cluster (or multi-stage) surveys. These more complicated forms of sampling do not, in general, produce a sample that has characteristics similar to the total population. However, when combined with appropriate statistical techniques, they produce valid (i.e., unbiased) estimates of characteristics of the total population. The value of these more complex survey designs is that they produce more reliable estimates of population characteristics (i.e., less variability) and are often cheaper to implement. The text, "Sampling Techniques" by William Cochran, provides an excellent (though somewhat mathematical) description of various types of survey designs.

Assessing and Accounting for Bias

Textbooks on survey methods typically assume that response rates will be 100%; i.e., all individuals selected for a survey will respond. In most real world settings, this is not the case. There will always be some number of individuals who will fail to respond for a variety of reasons. Response rates may vary by method of administration of the survey, inability to locate a consumer of services, anxiety in responding, or inability to understand the survey, to name just a few. Once a survey has any non-responders, no matter how few, the results of the survey are subject to two forms of bias: overt and hidden. Overt biases are those differences that can be observed between the individuals in the sample and those intended to be in the sample. Hidden biases are those biases that have not, or cannot, be observed.

The value of responses varies with how much you know about who does and does not respond. Even very low response rates can provide sound estimates of population responses if we know something about the characteristics about who responded and who did not. For this reason, it is extremely useful to collect **identifying information** as opposed to anonymous surveys for knowing who responded and demographic and (in the case of consumers) clinical information about respondents or identifying information that can be linked to such data.

Proportions of respondents in various demographic dimensions should also be inspected for departures from what would be expected in the client sample population, if that is known. It may also be possible to detect overt bias on the basis of theoretically critical issues, such as diagnosis or utilization. It is only possible to perform this investigation if there is some way to link survey responders to case records.

If you have information on consumer's demographics, diagnosis, and service utilization, you can compare sample characteristics with that of the population. If no significant differences

exist between the sample and the total population, then the data can be reported and there are no overt biases within the sample. If difference exist between the sample and the total population, then one will want to correct for overt biases. One method is to stratify the sample on those characteristics that differ between the sample and the total, then weight the sample using the total population's weights. Such a technique is known as a post-stratification of the sample. An example of this procedure can be found in Vermont's analysis of their consumer survey. Another way to accomplish this is to perform a *propensity analysis*. Simply stated, a propensity analysis is a statistical model that predicts whether an individual will be a responder or a non-responder. This prediction model is constructed using logistic regression, with response status (responder/non-responder) as the dependent variable. Then findings are compared for persons with similar propensity scores in different groups.

In all survey settings where there is even a small percentage of non-respondents, the potential exists for hidden biases. A hidden bias cannot be controlled for with statistical adjusting since the variable(s) that makes the respondent sample different than the non-respondent sample is not measured. Instead, one performs a *sensitivity analysis* to examine the impact of hidden bias. One way to do this, is to assume the most extreme values (both positive and negative) for the non-responders. This extreme setting is equivalent to saying that there is a large correlation between the outcomes that are reported and the variables associated with non-response. A second approach is to explore how related other characteristics would have to be to influence the results. In this setting, the analyst examines various scenarios for the hypothesized non-respondent data and calculates the correlation coefficients necessary to produce this data. In this way, one can examine the sensitivity of the mean of the respondent sample to non-respondent hidden bias.

Random vs. Stratified Sampling

When making decisions regarding how to sample, it is important to decide whether to sample randomly from a population or whether to divide the population into subgroups or strata. Though many issues need to be addressed in deciding on a sampling scheme (cost, precision, ability to implement), it is important to avoid what has been called "the tyranny of the large." Are there small subgroups of interest that conventional sampling won't allow us to find? For example, if one were to conduct a national survey, then people living in a frontier setting (<1 person per square mile) would be highly unlikely to appear in a simple random sample of 1,000 individuals. If this group were important to have in your sample, you should consider oversampling in this subgroup. Sometimes subgroups may be so small that it may be desirable or necessary to sample subgroups across certain providers or geographic boundaries.

Cross-sectional vs. Longitudinal Sampling

For measuring change over time, there are two broad categories of survey strategies: multiple cross sections and longitudinal designs. Within each of these categories, one can use any of the various sampling schemes mentioned so far, for example simple random samples or stratified samples. Multiple cross sectional samples (MCSSs) draw subjects from a group of people who meet certain criteria (e.g., received a service in last month). Each MCSS drawn over time might be composed of different persons. Longitudinal samples, conversely, identify a group of persons at baseline and follow these individuals over time. As a broad generalization, multiple cross sections are powerful methods for detecting system change, whereas, longitudinal designs are powerful methods for detecting individual changes.

A major advantage of **multiple cross sectional samples** is that they provide information about changes in population composition, characteristics, and attitudes over time. Another advantage of MCSSs is that if specific subgroups are continuously represented in a population over time, change in these subgroups can be tracked. Examples of subgroups are:

- One or more systems
- One or more plans or providers
- One or more age groups
- One or more diagnostic groups

A major disadvantage of multiple cross sectional sampling is that, in general, it does not allow us to track individual change. For example, we may learn that functioning is improving in a population, but we cannot be sure whether this is because persons are becoming higher functioning or because a population is adding more higher functioning persons.

In longitudinal samples, we follow a specified group of individuals across time. A major advantage of **longitudinal samples** is that such samples give us information about individual and subgroup change. These changes can be discriminated from changes in population membership. The major disadvantage of longitudinal sampling is that it is difficult and expensive to carry out. The difficulty comes from the difficulty in finding persons for longitudinal follow-up. This is referred to as respondent "retention" in discussions of evaluation methods.

Some ideas for increasing the rate of retaining respondents are:

- Obtaining the names of friends or relatives to contact if a respondent moves
- Establishing a toll free phone number so that persons who move can contact a project when the project can't contact them

Further ideas about retaining respondents are discussed in the following references:

Coen AS, Patrick DC & Shern DL. (1996). Minimizing attrition in longitudinal studies of special populations: an integrated management approach. *Evaluation and Program Planning*, 19:309-319

Ribisl KM, Walton MA, Mowbray CT, et al. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: review and recommendations. *Evaluation and Program Planning*, 19:1-25.

It is possible to combine cross sectional and longitudinal sampling. In these approaches, generally, the cross sectional sample is larger, while the longitudinal design follows smaller groups of individuals intensively. This combined approach is certainly the most comprehensive, but is also the most resource intensive.

We will end this section with a brief description of two other sampling approaches to consider: cluster sampling and inverse sampling.

- **Cluster, multi-stage sampling.**

A cluster or multi-stage sample implies sampling at more than one level. For example, we may first draw a sample of clinics, and then, within the sampled clinics, we may draw a sample of people. The advantage of this approach is that if one is doing face-to-face interviews, it reduces the number of clinics (or the number of interviewers) that need to be visited, and thus reduces the cost of the survey. The disadvantage is that the survey data requires special computer software or trained sampling statisticians to construct estimates and confidence intervals from this type of survey data.

- **Inverse Sampling**

An inverse sampling techniques is a power method to estimate the frequency of rare or unlikely events. In this sampling design, one keeps sampling until you see a pre-specified fixed number of rare events (e.g., suicide). This method allows greater precision for answering a specific question, but is not as good when one has many questions which need to be addressed.

Precision, Certainty and Sensitivity

In performance measurement, it is possible to design projects that are more or less precise for estimating the magnitude of measures, more or less certain with respect to whether groups differ, and differentially able to detect group differences as a function of their size. One key factor in determining these things is sample size.

If you are going to employ a sampling methodology, your MHSIP Report Card planning process will have to decide how precise, certain, and sensitive you want your measurements to be. Different policy uses require different levels of precision, certainty and sensitivity. For example, it is likely that measurements made for the purpose of quality improvement that will only be used internally require one level, whereas measurements made for monitoring contract compliance, that might be used in legal proceedings require another.

Precision and certainty are directly linked to sample size. Of course, if our goal is to give all consumers a voice in assessing the service system, then it would be ideal to collect data on everyone in a target population. However, this may be impossible for practical reasons. Or it may be possible, but only if some inexpensive method, such as mailed surveys, is used that may result in low response rates and increase the potential for both overt and hidden biases. Therefore, sometimes it may be better to seek a smaller sample from which you can obtain a higher response rate.

The advantages of **smaller samples** are that they are less difficult and costly to obtain. It is also easier to do the data management associated with smaller samples. Their disadvantages are that they are less precise, offer less certainty, and are likely to detect only large group differences. Smaller samples also limit the potential for subgroup analyses.

The advantages of **larger samples** are that they offer more precision and certainty, and are more likely to detect smaller group differences. They are also more likely to permit subgroup analyses. However, the larger a sample, the more difficult and expensive it is to collect data and more difficult to manage data (See Chapter 4, below).

Below are two quick rules for the relationship between sample size and precision. In general, there are two basic ways in which data are reported: the proportion of cases with a certain attribute (e.g., percent dissatisfied), or the sample mean (e.g., average score on the AIMS scale). In samples of about 100 persons the precision in estimating the proportion is within plus or minus 10 percent. This represents a so-called 95% confidence interval. Thus, if you observe that 32% of the sample is dissatisfied, you would say that you are 95% sure that between 22% and 42% of the total population was dissatisfied. If one had reported the mean, then the 95% confidence interval would be plus or minus .2 times the standard deviation. These methods work for the total sample as well as subgroups within the sample.

If one had samples of 400 per group of interest, the 95% confidence intervals for a proportion are within plus or minus 5 percent, and for the mean is .1 times the standard deviation. It should be noted, that to make the 95% confidence interval half the size (that is to go from plus or minus 10% to plus or minus 5%) we needed to increase the sample four-fold. This results is always true. So to make the 95% confidence intervals range plus or minus 2.5% would require, in general, samples of size 1,600.

The only exceptions to the above rules are due to the fact that in many instances, when we draw samples we have seen a large portion of the population that could be seen. This is known as the finite population phenomenon. To correct for this, one uses a finite population correction factor. Essentially, this correction factor implies that if you have seen a large percentage of a population, then your precision is greater than if you had only seen a small percentage. The formula for the finite population correction factor is the square root of $(1 - n/N)$, where n is the number of respondents in your sample, and N is the number in the total population. Using the above as an example, we have said, that if you have 100 respondents, then your precision is plus or minus 10%. Now if there were only 200 persons that could be observed, your precision would increase to plus or minus 7% (10% times square root of one half).

When designing a research study, investigators will often perform a power calculation. In the research setting, power is interpreted as the chance one would miss a statistically significant

difference due to the size of your sample. Thus, as the number of subjects in a study increases, so does the power of the study. In performance measurement systems, it may also be necessary to compare groups. The performance monitoring system differs from the research setting in that there may be many groups to compare, and more than one measure to compare. It is often the case, that we may wish to compare one program to another and we may wish to compare both programs to a common benchmark.

As an example, recall that above we noted with 100 respondents, that the proportion would be known to within plus or minus 10%. Thus if we were comparing a program with 100 respondents to a benchmark, if the program was more than 10% away from the benchmark, we would say that the program is significantly different (at the 0.05 level) than the benchmark. If we were to compare two different programs, each with 100 respondents, then if the programs differed by more than 14%, we would say that the two programs were statistically different than each other. As stated above, if we had 4 times as many individuals in each program, we would be able to say that differences of 7% were statistically different. These are statistically significant differences, not clinical or policy relevant differences. As the sample size in your data set grows, so will the number of statistically significant differences. It is important to decide how large a difference is necessary before one would become concerned. This discussion will help frame the question regarding how large a sample one needs to draw. The calculations above assume similar populations or adjustments for case mix differences have already been done. They also assume a single comparison is being made; multiple comparisons invite opportunistic chance findings and the interpretation of levels of significance is more difficult.

Some sources of information for doing power analyses are:

- Kraemer, Helena Chmure; Thiemann, Sue. (1987) *How Many Subjects and Statistical Power Analysis in Research*, SAGE Publications: Newbury Park, CA
- "A comprehensive list of power analysis software for microcomputers," found at www.insp.mx/dinf/stat_list.html
- *The Methodologist's Toolchest*, www.ideaworks.com
- SPSS Inc.
44 N. Michigan Ave
Chicago, IL 60611
Telephone: 312-329-2400
Fax: 312-329-3668

Two good general resources on sampling are:

- Henry GT (1990). *Practical sampling*. Newbury Park, CA: SAGE Publications
- Cochran, WG (1977). *Sampling Techniques* (3rd edition). New York: Wiley.

Reference Time for Measures

When measuring system performance, it is necessary to specify the time periods to which the measures should apply. These time periods are sometimes called "look back," or "reference" time periods. In discussions of evaluation methods, this topic is often discussed when instruments are described.

The options for this question can be stated in terms of shorter and longer reference periods. The reference period can be independent of the frequency and time period for data collection. For example, data could be collected annually, using a reference period of the past month for counting events and rating performance.

Frequently used reference time periods are:

- The last month
- The last three months
- The last six months
- The last 12 months

The options for this question can be stated in terms of shorter and longer look back periods.

Some advantages of **shorter reference** periods are that recall for survey questions is usually more accurate, and responses (e.g., satisfaction ratings) are more closely related to specific events or experiences. A major disadvantage of shorter reference periods is that responses are less representative of performance over time.

An advantage of a **longer reference** period is that, for a given level of accuracy, responses are more representative of performance over time. Some disadvantages of longer reference periods are that responses are less likely to be accurate and less likely to be tied to specific events or experiences.

How this trade-off between accuracy and representativeness is resolved in specific situations may depend upon whether large numbers of responses are anticipated within relevant subsets of the data; i.e., whether the respondents to the survey will represent the full range of experiences you wish to capture even if the reference period is short. If a large number of respondents is included within a sample being evaluated, and the time periods identified across respondents reflects the full range of interest, then using brief reference periods to optimize recall may be an appropriate strategy to obtain a good estimate. For example, you might be interested to ensure that survey responses reflect consumers' experience with hospital stays, which happen fairly infrequently. To accomplish this, you could either use a large sample size, thereby increasing the likelihood that adequate numbers of respondents will have been hospitalized recently; or you could use a relatively long reference period, thereby increasing the likelihood that any individual respondent would consider his or her hospital experience in answering the survey items.

While the Report Card does not specify a reference period for the MHSIP Consumer Survey, the most common reference period in use has been one year. This period is typically considered too long for valid recall of specific, frequently occurring events but may be appropriate for the sorts of global judgments solicited.

Timing and Frequency of Data Collection

Data may be collected at different time intervals.

This topic is discussed in evaluation materials in sections on procedures.

The possible options are:

- Monthly
- Quarterly
- Semi-annually

- Annually

There are several advantages of **shorter intervals**. These include greater retention and recall of intervening events for surveys, and the opportunity of identifying problems and improving services more rapidly.

The disadvantages are greater expense, and for surveys, greater burden and intrusiveness.

Data collected at **longer intervals** is less expensive, and in the case of survey data less burdensome and intrusive to collect.

Its disadvantages are that survey data is less likely to be accurate, it is less likely to be timely, and harder to interpret since it is less likely to reflect specific events. Also, if the look back time is set shorter than the data collection interval, a period of time will not be assessed.

As is true in general, the question regarding how often to collect data is a balance between cost and intrusiveness on the one hand, and precision and timeliness on the other. Policy makers and those who will make use of the consumer surveys will need to weigh the various advantages and disadvantages of the timeframe of the surveys.

The Task Force Report calls for the following time frames for consumer reported data:

- Consumer Survey: at discharge for all service recipients; annually for a representative sample of service recipients
- Consumer report items (including items from SF-36, Rosenberg's self-esteem scale, symptom distress scale, etc.): at admission, three months after treatment begins (or end of treatment) and a year from admission for those still receiving services.

To date, most MHSIP Report Card projects have collected Consumer Survey data annually. We are not aware of any projects that have collected data as often as quarterly, as is suggested by the Task Force Report.

Versions of the MHSIP Consumer Survey

The MHSIP Task Force report issued in April 1996 included a new consumer survey developed for the purposes of the Report Card. This survey contains 40 items, and has since been field tested extensively. Results of these tests have led to certain modifications to the survey. In addition, most states that have adopted the MHSIP Consumer Survey have made slight modifications to its wording prior to fielding it. Finally, there has been widespread interest in a shortened version of the MHSIP Consumer Survey. These factors have led to a somewhat confused situation regarding the survey, with several different versions in use across the country all being referred to as "the MHSIP Consumer Survey." Since you will have to choose which version of the survey to field, we will describe the most prominent versions currently in use.

The most notable and well accepted change that has been made to the original survey is the modification of what have been referred to as the negatively worded items, such as "I was unable to get some services I wanted because I couldn't pay for them" and "Staff were not sensitive to my cultural/ethnic background." Seven items were constructed in the negative to try to solicit feelings of "dissatisfaction" as distinct from satisfaction. Strong evidence has accumulated that instead of dissatisfaction, these items appear to measure an unwanted methodological dimension. This evidence comes both from statistical analyses (factor analytic and structural equation modeling) as well as from reports from respondents that they found the items confusing. Accordingly, they have been revised to the positive phrasing, consistent with the other items in the survey; e.g., "Staff were **not** sensitive to my cultural/ethnic

background," has been revised to, "Staff were sensitive to my cultural/ethnic background." Although this solves the short-term problem of "bad items", it is not consistent with the original intent to measure dissatisfaction as distinct from satisfaction. Research is currently being conducted on this important issue, but at this time there is not a definitive answer.

The MHSIP Consumer Survey was constructed to measure four major domains: Access, Appropriateness, Outcomes, and General Satisfaction. Using confirmatory structural equation modeling techniques, researchers in Colorado analyzed data from five states. Their analyses confirmed these domains and suggested that it was possible to reduce the number of items below 40 and still have reliable scales for each major domain.

In response to strong interest in a short version of the MHSIP Consumer Survey, researchers in Colorado and Texas began work on a 21-item version of the survey specifically tailored to the needs of NCQA and their primarily outpatient provider network of health plans. This short version is based on structural equation modeling of the data used to confirm the instrument's domain structure. This analysis confirmed that the 21-item version adequately represented the major domains of access, appropriateness, satisfaction, and outcomes, and that using the "short" version did not change the relationships among variables. That is, the four factor structure of the 40-item survey was the basis for developing a shorter version; i.e., certain items were selected from each of the four factors to reduce the total number of items but retain the underlying constructs of access, appropriateness, satisfaction and outcomes. The short version was developed specifically for NCQA in considering the addition of a behavioral health consumer survey to HEDIS, its performance reporting requirement for NCQA accreditation. To meet NCQA's needs for a broadly applicable survey that could be used to evaluate health insurance plans, certain questions were included or excluded without respect to their loading on the factors. In addition, some wording of items was changed to make the survey more appropriate for NCQA purposes; specifically, the term "agency" was replaced by "health plan," and "staff" was eliminated in some cases by rewording the item, and replaced in a few instances by "caregivers."

The 21-item version, while still fairly new, does appear to successfully measure consumers experiences in the areas of access, appropriateness, satisfaction and outcomes. It is still in the process of being tested more extensively to confirm its appropriateness for a wide range of mental health consumers, programs, and health plans.

Given this, how should you decide which version of the MHSIP Consumer Survey to use? Some general guidance about instruments and their properties may be helpful.

Performance domains usually consist of more than one measure. Similarly, surveys and scales constructed from survey items can have fewer or more items. There are technical reasons to prefer more measures in an area. However, because of resource and burden concerns, people usually look for ways to abbreviate data items. Abbreviated measures of indicators and surveys are often referred to as "short forms."

This topic is discussed in evaluation materials in the sections on instruments.

There are various scientific properties that a measure must have to be consistently accurate. These are known as psychometric properties. We discuss some of these further below. Psychometric properties are a significant factor when selecting an instrument. While psychometric properties are not necessarily directly related to length, tradeoffs often have to be made between better psychometrics and brevity.

One major advantage of employing **more questions** or items is that indicators and scales constructed from more items tend to be more reliable or consistent. In addition, specific items may be valuable in and of themselves in addition to their contribution to a factor or subscale. Major disadvantages are that larger numbers of items or questions are more burdensome and require more resources to process.

A major advantage of **fewer questions** or items is that their burden and resource requirements are less. A major disadvantage is the likelihood of lower reliability.

While the short version does appear to demonstrate similar psychometric properties to the full 40-item version, we recommend using the full version of the survey, unless resource constraints are severe as to preclude the longer survey. The major reason for this is that each item was developed for the survey because it was judged important in and of itself and for some of these items, stakeholders may want to analyze data at the item level and not just the subscale or factor level. If you decide to use a shorter version, we recommend that you begin with the 21-item survey developed for NCQA and consider adding additional questions from the original set of 40 items that you may feel is important to analyze separately.

In some cases, MHSIP users have changed the response choices, typically reducing the number of options from five to three or two. We do not recommend changing the **response options** of the MHSIP Consumer Survey items. If persons have difficulty responding to the number of survey responses, do not, for example, change them to yes or no. You can however, break-up the responding, for example, by first asking do you mainly agree or disagree or are you neutral, and then, in the case of a person who agrees, do you strongly agree or just agree and so on. We recommend not changing response categories, because doing so will make your data less sensitive to individual differences and make it more difficult to compare with the data from other MHSIP projects.

Method of survey administration

If you elect to use one or more surveys, you will need to decide how to administer the surveys. There are two very important ways in which survey data are influenced by method of administration. These are:

- The number of persons who respond
- The extent to which persons express their true beliefs or feelings (bias)

In evaluation materials, different methods of administering surveys are discussed in materials on survey research and in sections dealing with evaluation procedures.

The most commonly considered methods of survey administration are:

- Mail surveys
- Phone surveys
- Point of service surveys
- In-person interviews
- Combinations of the above

There are other issues in survey administration particularly related to methods for obtaining more responses. These are:

- Methods for obtaining consent to be contacted
- The use of inducements and incentives for participation
- The use of identifying information for locating persons and for keeping track of who does and does not respond

Mailed and self-administered

Mailed surveys have the advantage of potentially reaching the largest number of possible respondents. While some people you may want to survey may not have phones and many may not be coming in for service, most will have mailing addresses.

Another advantage of mail surveys is that unless someone's mail is opened by another person, the request for participation is private. Also, the respondent can choose a private time and place to respond. Another advantage of mailed surveys is that if cash or other incentives are being used, they can be contained in the envelope.

Finally, mailed surveys are relatively inexpensive to administer. However, since the response rate for surveys can be low, the cost per completed survey can be higher than it might first appear, especially if multiple mailings and telephone follow-up are employed.

The disadvantages of mailed surveys are poor response rates. Typically, response rates are between 20 and 30%. An additional disadvantage is that mailed surveys can only go to persons with addresses. You may also experience difficulties if you want to survey persons who reside in treatment facilities. These persons may not have private and confidential mail access.

Mail out methods can differ in their strategies for increasing response rates and/or accounting for biases due to non-response. The following list describes some features different methods may have:

- Anonymous versus identifying information: identifying information permits follow-up for non-respondents and adjustments for overt bias as described above
- Parallel postcard to be returned if survey returned: This strategy calls for enclosing a postcard along with the survey and asking respondents to return the postcard when they return the completed survey; the postcard contains identifying information but the survey itself does not. This allows for tracking and adjustments while maintaining strict anonymity of survey responses.
- Use of postal permit to reduce cost
- Non-identified envelopes to protect confidentiality of service use (i.e., envelopes containing surveys do not have "Department of Mental Health" return address)

A full discussion of these and other options can be found in the following references:

Salant, P. & Dillman, DA., (1994) *How to conduct your own survey*. London: Wiley.

Dillman, DA (1978). *Mail and telephone surveys: the total design method*. New York: Wiley.

Dillman, DA (1991). *Mail surveys: a comprehensive bibliography, 1974-1989*. Chicago, Ill.: Council of Planning Librarians.

Telephone Surveys

One advantage of **telephone surveys** is that they can result in higher response rates. Another is that this method requires no reading ability and the persons making the survey call can answer simple questions regarding its completion.

Several disadvantages must be weighed against the advantages. First, not all potential respondents will have telephone numbers. Second, telephone recruitment can violate privacy unless a protocol is developed for the situation in which someone other than the respondent answers the phone. Third, the respondent may not be able to find a private space in which to respond and the respondent may view the interviewer as violating their privacy. Finally, phone interviews are more expensive than simple mail-outs, especially when repeated call backs are necessary as they often are.

Telephone interviews, like in-person interviews, discussed below, also have the disadvantage of requiring interviewer training. Just a few of the areas in which interviewers must be trained are:

- Recruiting respondents
- Avoiding biasing responses
- Accurately recording responses
- Reassuring respondents who are upset by questions

Administration at Point of Service

One advantage of **point of service** interviews is a high response rate, particularly if the respondent is asked to complete the survey before leaving the point of service. Another is that the person distributing the survey can answer simple questions regarding its completion.

There are several disadvantages. One is that a point of service survey only reaches persons who appear for service. This can be partially addressed by using a wide window for data collection. A second is the method requires someone to distribute and receive the survey. There is consensus that this should not be a service provider or anyone associated with service provision, particularly when it comes to receiving surveys. The effect of this might be to bias persons to give favorable responses or worry about the treatment they receive. One option for distributing surveys would be a clerical person. Another would be a person involved with quality monitoring. To repeat, the most ideal person would be someone viewed as outside the service system.

In-person Interviews

In-person interviews have the advantage of having the highest response rate. This method also requires no reading ability and allows for persons to be helped to complete the survey; i.e., interviewers or surveyors can play a more or less active role in facilitating survey responses depending on the individual needs of a respondent. In-person interviews also allow for interviewer ratings. For example, with proper training, an interviewer can complete a rating of consumer level of functioning or symptomatology if this measure is also included as a performance measure. Short of this, interviewers can provide basic information that may shed light on whether the respondent appeared to understand the survey process.

In-person interviews have the disadvantage of being the most expensive method. Like the telephone interview, they require resources for training interviewers as well as survey administration. It is also the most intrusive, and possibly the most burdensome method. Additionally, in-person interviews require careful consideration to issues such as interviewer safety, where interviews will take place, and respondents' comfort levels with persons of the opposite sex. All of these specific examples have emerged as difficult issues in one or more performance measurement project.

Combined Methods and Field Experience

Combined methods can be used to minimize some disadvantages of single methods. However, they also combine certain disadvantages such as risks to privacy and expense. An example of common combined method is following a mail-out of the survey with phone calls to non-respondents attempting to complete interviews over the phone.

Most MHSIP efforts to date have used mailed surveys. At least one project used consumer interviewers and one used point of service administration. To date, no single method of administration has become the standard, but rather individual projects have sought to weigh the advantages and disadvantages of the various choices in a local context.

Response rates and cost effectiveness (i.e., cost per completed survey) have been two of the major forces driving selection of administration method. Table III-1 below, lists different methods of administration and reports their return rates found in Delaware's pilot implementation of a consumer survey based in part on the MHSIP Consumer Survey.

Table III-1

Methods of survey administration, completion rates, and estimated costs per completed survey: Results from a pilot study in Delaware

<u>Method of administration</u>	<u>Completion rate (%)</u>	<u>Estimated cost per completed survey</u>
Mail surveys	26.7	\$29
Provider distributed	74.0	\$5
Individual interviews	90.0	\$26
Small group sessions proctored by consumer interviewer	70.0	\$40

Cite: Tippet, Maurice; presentation at 47th National Conference on Mental Health Statistics

Obtaining Consent

Successfully collecting survey data from consumers identified by the sampling strategy chosen is necessary to ensure the desired precision for measures. For this reason it is important to consider how to solicit consumers to participate, how to conduct the informed consent process, and what incentives and inducements can be used to increase participation.

It is worth stressing that despite the technical reasons that it is desirable to collect data from all consumers targeted, each individual should be approached in a non-coercive way and guided through an informed consent process. Ultimately, it is up to the individual to decide whether or not to complete the survey; it is your job to ensure that the consumer is provided with all the appropriate information so that he or she can make a fully informed choice.

Here are a few suggestions about obtaining consent:

- Make sure the consumer understands what the study is about, and what they are being asked to do.
- Explain the importance of the survey and what it is trying to find out.
- Describe how data might help to improve services.
- Answer any questions that he or she might have about the study.
- Review your procedures regarding privacy, anonymity and/or confidentiality.
- Offer the report and/or opportunity to attend meetings to discuss results at end of study.

Several different procedures for **obtaining consent** have been used in different settings. These include:

- Contact by a QI person or ombudsperson to obtain consent followed by contact by interviewer or data collector
- Consent process and data collection conducted by interviewer
- Inclusion of consent form in mailing containing survey
- Initial contact by consumer organization followed by referral to interviewer
- Consent and data collection conducted by a peer

No one process for obtaining consent has shown itself to be preferable to all others. In devising your consent process it is important to involve stakeholders, particularly consumers, and to consult requirements or regulations as a result of approval by an IRB.

Inducements and Incentives for Participation

Without being coercive, there are strategies that you can employ to increase participation from consumers. We refer to these strategies here as inducements (i.e., actions that might lend credibility to the project and thus raising the appeal of participation), and incentives (i.e., actions that provide more direct compensation for participation).

Examples of **inducements** include letters supporting the project from important stakeholder groups. Texas and Maryland used letters in this way. Copies of their letters are contained in Appendix B.

Incentives can be cash, coupons or other means of compensation. Some have suggested that cash or money orders are preferable to checks, since many respondents may not have bank accounts.

Surveyors/data collectors/interviewers

If you decide to implement a point of service, phone, or in-person survey, you will have to decide who will distribute and administer the survey. Whom you choose to do so may affect your response rates and the quality of your data.

This topic is often discussed in the procedures section of evaluation materials.

The most frequent options for survey distribution and/or administration are:

- Providers
- Consumers
- Third parties

Data collection by **providers** often has the advantage of being convenient. They are often in contact with survey respondents for other reasons, they know the respondents and may be trusted by them, and they may have interviewing skills.

However, data collection by providers has the very important disadvantage that persons may be reluctant to be completely honest with them because they are part of the system being assessed. Consumers have expressed fear of retaliation in some cases and more general feelings of vulnerability when providers are involved in collecting evaluative information. It can also take away from provider time with consumers that should be spent on service provision. This may result in low response rates and/or spurious responding. In addition, consumers may not feel free to choose to decline to participate when asked by providers. *All in all, the disadvantages of using providers for data collection from consumers clearly outweigh the advantages in all but exceptional cases. Thus, we strongly recommend separating the data collection process from service providers and the service provision process.*

There is some opinion that survey distribution and administration by **consumers** will cause a greater percentage of consumers to participate in surveys and to be more forthcoming with their views.

- Given the principle that the method is the message, it may also be that using consumer interviewers adds empowerment value to process.

The disadvantage to this may be that consumers usually have less experience than other persons who might be employed to conduct interviews; consequently, training needs may be greater, particularly when compared to the option of using a professional survey research firm.

A literature review on experience with employing consumers or peers as interviewers or other types of data collectors prepared by the New Mexico Division of Mental Health is included in Appendix C.

To date, at least one state, New Mexico, has used consumers to administer a MHSIP survey. A summary of the New Mexico experience follows.

Case Study: Consumers as Surveyors in New Mexico

Consumer surveyors were recruited by word of mouth and mailed notices to provider agencies, advocacy groups, and self help organizations. A simple application form was required and virtually all applicants were invited to participate in one of a series of one-day regional trainings.

The training was designed to familiarize consumer surveyors with the MHSIP initiative, to provide them guidelines/protocols in how to conduct the surveys, maintaining safety, responding to crisis, etc. Orientation to the administrative details of processing forms, billing etc., was also provided. The training sessions also provided opportunities for surveyors to practice in role play and problem solving exercises. A copy of the training manual is on file at HSRI.

Surveyors were paid a flat rate of \$20 per completed interview; consumers completing the interview were not compensated for their participation. Surveyors were also provided necessary tools and resources including such things as papers, pre-stamped envelopes, and even staplers. Pre-paid long distance phone cards were also provided. In some very rural areas requiring significant travel, supplemental allowances for these expenses were also paid.

The survey protocol required a face-to-face meeting in a public setting. Surveyors made individual arrangements with consumers to conduct the surveys. In some cases where a fact-to-face meeting was not possible, the survey was conducted by telephone at the discretion of the surveyor.

A meeting with the surveyors was convened near the end of the project in order to debrief and learn from the effort. Overall, the surveyors felt that their work had been a very worthwhile experience which had improved their own sense of confidence and self esteem. The consensus amongst them was that they had been fairly compensated and would be interested in future opportunities to work as a surveyor.

The surveyors reported that in many cases it was only necessary to present the consumer participant with the survey forms which were then self completed. In other cases consumers required considerable assistance including having the entire survey read item by item.

There was general consensus amongst the surveyors that having consumer peers as surveyors was very valuable. The empathy, understanding and ready help substantially improved the consumers' willingness to participate and response rates along with the consumer's overall experience of being surveyed.

Follow-up quality checks were conducted on a random sample basis. Irregularities and questionable practice in completing and submitting surveys was found with only one out of over thirty surveyors.

In summary, the New Mexico experience demonstrated that employing consumers as peer surveyors is a feasible if not optimal approach to collecting consumer survey data.

One advantage of **third parties** is that they may not cause respondents to alter their responses for fear of consequences since they are likely to be perceived as separate from the mental health system. The types of third parties you can employ range from students looking for part-time work to a professional survey research firm. A professional firm will likely bring a great deal of experience and infrastructure for collecting data to the project, but can also be prohibitively expensive. Experience suggests you can probably find a group of persons with appropriate skills and experience in the labor market.

A disadvantage of third party interviewers may be that they do not have credibility with some consumers. Some consumers may decline to be interviewed or provide data that is spurious. More generally, persons who do not have experience with the mental health system or with persons with mental illness will need to be trained extensively about these issues, and even so may not be as sensitive as is desired.

Translations

If you wish to survey non-English speaking persons, you will have to use translations of the MHSIP Consumer Survey. The reason for using translations follows from issues already reviewed. If you need to be able to generalize conclusions about consumers' views of services to these populations, you must include them in your sample.

A number of states have developed translations of their versions of the MHSIP instrument. At least a few Spanish translations have been developed and one state translated the instrument

into Vietnamese. We include a few translations in Appendix D. When additional translations become available, information about them will be posted on the MHSIP Report Card website, www.mhsip.org. If you need to include a population in your survey for which no appropriate translation yet exists and you are interested in developing one, it will be important to follow certain methodological procedures to ensure that the resulting instrument yields valid and reliable data. Such procedures require the investment of effort and resources, and it is strongly suggested that you have the appropriate expertise within or available to the group that is to do the work. This section reviews some important decisions you will have to make about how you develop your translations.

The obvious advantage of **translating** your consumer survey is the ability to collect data from a group you deem important. As cultural and linguistic diversity continues to grow in nearly all parts of the country, and along with it a greater emphasis on cultural competency of the mental health system, being able to obtain the opinions of non-English speaking consumers will become more important.

The disadvantage is that this may add expense and effort to your project, especially if you are using telephone or in-person interviews. In these instances you will not only have to develop a translated survey, but also will need to hire bilingual interviewers.

You may also wish to use interviewers from the ethnic and cultural groups providing data. The advantage of this is that respondents may be more open with interviewers from the subgroups to which they belong. The disadvantage is that it may be difficult to find interviewers from certain subgroups, and some persons may find this type of "ethnic matching" inappropriate.

Issues in Developing Translations

In reviewing the lessons learned from case studies of a variety of approaches to developing translations, McKay, et al (1996) suggested a framework for the major issues that must be addressed. These include selecting a translation method, identifying the objectives of the translation, obtaining and training translators, managing the translation process, pretesting the translated instruments, and training interviewers. Each of these issues is discussed in turn below.

Selecting a translation method

Most translations are done using either *direct translation* or *back-translation*. Direct translation is simply a one-way translation of an instrument from one language to another. With back-translation, a separate translator or team translate the translated material back into the original language. This process can reveal possible weaknesses in the translation, which can then be used to achieve a version that is more faithful to the source instrument. Although back-translation is normally recommended, the process does require greater investment of resources. McKay et al. (1996) note that, if other methodological issues are addressed appropriately, a direct translation may be quite adequate.

Because the process of translation can also reveal difficulties in the body of the source instrument, and if modification of the source instrument is permissible, a third method can sometimes be used. *Decentering* entails successive changes in both the source and target-language versions until they are both comparable and satisfactory. Because of the value of using comparable instruments across states, decentering in any specific survey project is not recommended. However, if problems with the source instruments identified during back-translation are noted and shared, future standardized versions of the MHSIP Consumer Survey can incorporate the benefits of this method.

Identifying the Objectives of the Translation

Before translators begin, it is important that translation objectives be clear. A *literal translation* is one in which the wording in the target language is the dictionary equivalent of the source. Because literal translations cannot allow for the dynamic nature of language expressed in regional and subcultural adaptations, they often result in stilted, nonsensical or even incomprehensible questions. A *conceptual translation* allows the translator more flexibility in choice of wording in the target language in order to reflect the connotations in the source. McKay et al. (1997) strongly recommend that a conceptual translation be the minimal objective. A third possible objective, requiring still more careful refinement, is to achieve a *culturally equivalent translation*, in which shadings of meaning and thought patterns in the target language are cultural approximations of those in the source. Attaining full cultural equivalence of the current form of the MHSIP Consumer Survey for some population groups may not be possible, since such included concepts that are included in the survey such as empowerment, do not have similar meaning in all cultures.

Selecting and Training Translators

Translators should be chosen on the basis of their ability to provide a conceptual translation. This requires that they be fluent in both languages and familiar with the particular cultural group for which the translation is being prepared. Even though many different population subgroups speak Spanish, for example, there are important cultural differences between Hispanic or Latino populations that are reflected in language usage. Translators should also be flexible, so that they can adapt their work to the requirements of the survey. They should become very familiar with the purpose and wording of the English-language MHSIP Consumer Survey so that they understand clearly the concepts that underlie each of the questions. Training should include a careful review of the knowledge, beliefs, and values held by the target population that are associated with the content of the survey. Reading level, vocabulary, and cognitive capacities of respondents should be anticipated. Translators should also be familiar with the customs for social encounters and discourse that are relevant both to the way the interview is conducted and to the way that questions may acceptably be asked.

Managing the Translation Process

The translation process may include several steps that depend upon which methods and objectives have been chosen. In one possible scenario, a lead translator creates an initial translation, which an editing team then corrects or refines. If back-translation is used, a second team may create an English-language version, and then both teams may come together to resolve identified discrepancies. Pretesting, described below, may reveal problems in the field that call for repeating the process on some portion of the survey.

Pre-testing the Translated Instruments

Preliminary versions of translated instruments should be carefully evaluated before release for routine use by pretesting them with members of the target population. Interviewers are a good source of information about the ease of use of the instrument in an interview setting and about the apparent experience of interviewees. More rigorous approaches to evaluating a translation are derived from research on the *cognitive aspects* of interview situations. These include *debriefing* both interviewers and respondents after the interview to learn from their experience of dealing with the questions, and interviews where subjects "*think aloud*" to reveal how they interpret the questions. The findings from these pretests are fed back to the translators; revised versions are likewise subjected to pretesting.

Training Interviewers

Training for interviewers is similar to the preparation of translators. They should be trained on the English-language version so that they are very familiar with the purpose of each question on the survey. They should be well grounded in the cultural world of the respondents, and they should be very familiar with culturally appropriate conduct in social situations.

Whenever possible you should begin with an already developed translation developed in an acceptable manner (e.g., backwards translation). Remember, however, that a language may have dialects or idioms unique to a particular region. Thus, if you start with an existing translation, you should test it in your locality to make sure there are no differences in dialect or usage.

III. Non-Survey Indicators Based on Administrative Data

As noted earlier, the MHSIP Report Card relies on several types of data. Here we discuss the design and methodological issues related to using administrative data, which can include enrollment and encounter/claims data, case records, and other records that are routinely kept by a service provider or management organization.

While the basic framework for each indicator is included in the Task Force Report, there are many technical details about data collection and data application that are not specified. This leaves numerous questions about how best to operationalize many of the indicators and measures and allows for multiple interpretations and strategies for full implementation of the MHSIP Report Card.

While much progress has been made in resolving these questions for those indicators using the 40-item MHSIP Consumer Survey as the data source, the indicators and measures requiring other sources of data (e.g., administrative and fiscal data, encounter data, clinical data, etc.) have not been well developed. Until there is either broad consensus on how to operationalize these indicators and measures, or clear direction from the MHSIP Task Force or other authority, it is likely that there will be considerable differences in interpretation and implementation of these elements of the Report Card. Even with more detailed guidance from the MHSIP Task Force, it is likely that local idiosyncrasies in administrative data systems will require Report Card implementers to make their own judgements in certain cases.

Accordingly, for organizations that prepare a Report Card using the "non-survey indicators," caution should be exercised. Before embarking on such an effort, careful attention must be paid to a host of methodological issues ranging from data collection to definition of data elements and data presentation. Before making any comparisons across or between systems, careful examination of the comparability of the definition of data elements, data sources, data collection, case mix adjustment and sampling strategies, as well as actual methods of calculation is essential.

In many instances the data may not be readily available or not available at all. In some settings, it may simply not be feasible to utilize some of the indicators. Data requirements may exceed the capacity of existing information systems and administrative infrastructure or may create an insupportable burden for clinicians, administrators, and support staff.

In addition, it is important to recognize that, for virtually all of the indicators, there are no established benchmarks. While the concepts and concerns supporting each indicator may have some face validity and value, there is no established point of reference for the indicators that might begin to describe average, superior or sub-optimal performance.

Indicator Access 1 provides a good example of these issues. The indicator is expressed as a calculation in which there is typically a numerator and a denominator. The indicator is then expressed as a quotient or rate derived from the specified calculation.

The average length of time from request for services to the first face-to-face meeting with a behavioral health professional.

The indicator reflects an easy to understand concern about the need for quick easy and convenient entry into services.

The measure for the indicator is :

Numerator

The total time between request for services and the first face-to-face contact with a licensed mental health professional for adults (or licensed behavioral health professional for children and adolescents) for new admissions during the reporting year

Denominator

The total number of new admissions.

While the concept is straightforward, implementation is not. Each element, such as *request for service*, *face-to-face contact*, or *new admission* lacks a precise definition. These elements do not have universal definitions and are potentially subject to a wide range of interpretations from setting to setting. Even the terms *adult* and *child*, which might seem simple and straightforward, are applied differently in different settings and need to refer to specified ages and cut-off points for each category.

Complete specification and definition of terms developed in New Mexico for Indicator Access 1 from the Report Card is found in Appendix E.

In addition, collecting data for this indicator and measure could prove to be very problematic for many providers and may even require the introduction of new data systems and data infrastructure. How and where is the time of the first request for service recorded? Is it automated? How is the time of the first face-to-face contact recorded and how is it linked to the time of the request? Are encounters coded such that brief triage contacts can be distinguished from more thorough evaluations of consumer needs? If a consumer misses or cancels the scheduled appointment time, how is this captured and how is it figured into the average time between request and first contact? Implementation of this indicator will ultimately require answers to these and many other questions.

In summary, the following guidelines, suggestions, questions, and caveats should be considered in implementing and operationalizing indicators derived from administrative data sources:

1. Operational definitions should be provided for all of the terms used. For example:
 - the criteria applied to the term "serious emotional disturbance" should be specified along with the DSM-IV diagnostic codes used in a client/clinical data base
 - qualitative descriptions of services should be elaborated to distinguish between categories and types of services such as emergent and non-emergent situations or triage evaluation and comprehensive assessment
 - the corresponding provider or other service codes (e.g. CPT codes) included in an encounter/service data base should be specified
 - terms, such as "consumer-run" or "natural setting," may be difficult to define or to use in specifying required data contained in an administrative databases.

1. Data and database issues need to be addressed. For example:

- from which database(s) will the data be extracted; e.g., Medicaid claims, client registration data, telephone logs, etc.?
- if the Report Card allows for different data options, (e.g. consumer self-report vs. administrative records) how will the data source be selected?
- are all the necessary data in one database? any database?
- can one data base be linked to another to include all of the elements necessary to complete a measure?
- does the database contain the information necessary to distinguish among distinct groups and special populations for whom the measure will be computed; e.g., diagnostic groups, age groups, those seeking emergent services, etc.?
- are there sufficient resources to extract or gather this information and compute the indicator?

III. Clinician Rated Data

The major clinician reported data for the MHSIP Report Card are the Abnormal Involuntary Movement Scale (AIMS) to assess negative effects of neuroleptic medication, the Clinician Alcohol and Drug Use Scales, and the Child and Adolescent Functional Assessment Scale. Here we discuss the design and methodological issues related to collecting these or other clinician reported data including sampling, reference time for measures, frequency and timing of data collection, and training and assessing reliability.

Sampling

If clinician ratings are not being collected routinely, you may want to limit this data collection to a sample of consumers. The sampling considerations and procedures described above for the consumer reported data are similarly applicable in sampling for clinician rated data. In fact, it may be desirable to draw a single sample for consumer reported data and clinician rated data so that these different indicators can be compared for the same persons.

Reference Time for Measures, Frequency and Timing of Data Collection

Once again, reference time or look back time and frequency and timing of data collection are equally applicable to clinician rated data as consumer reported data; refer to the discussion above.

There may be a few unique issues for clinician data. The reference time may already be attached to some clinician rated instruments (e.g., a level of functioning scale may specify a rating of functioning over the last two weeks); if so, you should consider carefully whether it makes sense to revise the time frame. Such a revision may prevent comparing the resulting data with statistical norms or may affect the instrument's psychometric properties in ways it is impossible to predict. You will have to weigh these considerations with the desired time frame for your performance measurement system.

The MHSIP Report Card provides recommendations for the frequency and timing of data collection for clinician ratings. It suggests that the clinician assessments be collected along with the consumer self-report items; i.e., at admission, three months after treatment begins (or at end of treatment), and a year from admission for those still receiving services.

Establish and Monitor Reliability

The most important methodological consideration when using clinicians' assessments in a performance measurement system is demonstrating that clinicians are reliable raters. There are two primary steps in this process: training raters and assessing the reliability of ratings.

Train raters

Before collecting data from clinicians for your performance measurement system, you should train those clinicians who will be completing assessments in the content of the instrument(s) and the procedures of the data collection. In some cases, the instrument(s) to be used in the Performance Measurement is already routinely collected from clinicians and entered into the MIS system. Despite clinicians familiarity with the instruments in cases like these, it is valuable to review the instructions for completing the instrument(s) before using these data for Performance Measurement purposes.

Many clinician rated instruments have accompanying training materials, such as case vignettes and videotapes, available—these may be good, economical options for a public system. You may choose to bring in the developer of the instrument or another expert to conduct a training. Finally, given staff turnover, it may be wise to periodically train clinician raters, or to include training on instrument completion as part of orientation to new clinicians. Other suggestions include:

- Use train the trainers approach
- Certify assessors
- Have an ongoing mechanism to train new staff

Assess Reliability

Once training is complete, you will want to assess reliability to ensure that it meets standards of acceptability. We also recommend that if you are collecting data from clinicians over time, you should periodically assess inter-rater reliability. The procedures for assessing reliability are detailed in Chapter 4.

Chapter 3: Recommendations

- It is extremely useful to track identifying information as opposed to anonymous surveys to allow you to account for bias caused by non-response
- If you wish to compare groups you need to have larger group sample sizes than if you only wish to estimate the magnitude of a measure for one group
- We recommend not changing response categories, because doing so will make your data less sensitive to individual differences and make it more difficult to compare with the data from other MHSIP projects
- It is important to consider how to solicit consumers to participate, how to conduct the informed consent process, and what incentives and inducements can be used to increase participation
- It is desirable to collect data from all consumers targeted; each individual should be approached in a non-coersive way and guided through an informed consent process
- We strongly recommend separating the data collection process from service providers and the service provision process
- It is strongly suggested that you have the appropriate expertise within or available to the group that is to do translation work of the MHSIP instrument
- Back translation is normally recommended; the process does require greater investment of resources
- Interviewers should be well grounded in the cultural world of the respondents, and they should be very familiar with culturally appropriate conduct in social situations
- Before making any comparisons across or between systems, careful examination of the comparability of the definition of data elements, data sources, data collection, case mix

adjustment and sampling strategies, as well as actual methods of calculation is essential

- When using clinicians' assessments in a performance measurement system it is important to demonstrate that clinicians are reliable raters

Reference: McKay RB, Breslow MJ, Sangster RL, Gabbard SM, Reynolds FW, Nakamoto JM, Tarnai J (1996). Translating survey questionnaires: Lessons learned. In Marc T. Braverman, Jana Kay Slater (Eds.), *Advances in Survey Research*. New Directions for Evaluation, no. 70, Lois-ellen G. Datta (Editor-in-Chief). San Francisco: Jossey-Bass.

Center for Mental Health Services (1998) *Methodological Standards for Outcomes Assessment*