Design Options for Home Visiting Evaluation

MEASUREMENT BRIEF

Selecting Data Collection Measures for MIECHV Benchmarks

May 2011

Introduction

Home visiting programs seek an array of goals such as improving maternal and child health, parenting practices, school readiness, and the prevention of child abuse and neglect. The intent of this document is to support the Maternal, Infant and Early Childhood Home Visiting (MIECHV) program as part of the provision of technical assistance to funded grantees. Specifically, this brief focuses on the selection and development of performance measures or indicators to achieve those goals related to the legislatively mandated benchmark areas.[1]

The published requirements for the program state that grantees supported with MIECHV funds must collect quantifiable data for all eligible families enrolled in the program across six benchmark areas.[2] MIECHV Program grantees must specify in their Updated State Plan submitted to the Department of Health and Human Services an indicator for each individual construct listed under the benchmark areas in the guidance document or Supplemental Information Request (SIR) issued on February 8, 2011. Grantees also need to define what constitutes improvement over time for each indicator and propose a plan to collect and analyze the relevant data in order to report on the selected indicators annually.

The legislatively-mandated benchmark areas are:

- Improved maternal and newborn health;
- Prevention of child injuries, child abuse, neglect, or maltreatment, and reduction of emergency department visits;
- Improvement in school readiness and achievement;
- Reduction in crime or domestic violence;
- Improvements in family economic self-sufficiency; and,
- Improvements in the coordination and referrals for other community resources and supports.

Documenting improvement within each of these benchmark areas is a legislatively mandated requirement for the program. Assessing progress towards program goals -- both for accountability but also for continuous quality improvement purposes-- involves a multi-step process of indicator selection and development that begins with identifying and clearly

[1] For purposes of the MIECHV program, the terms “indicator”, “performance indicator”, “measure”, and “performance measure” are used as synonyms in the text. They should be distinguished from “measurement tool” or “measurement scale”. An indicator utilized to track quantifiable improvement under the MIECHV program may or may not require the use of a measurement tool.

[2] The “Supplemental Information Request for the Submission of the Updated State Plan for a State Home Visiting Program” (SIR) full report is available at url: http://www.hrsa.gov/grants/manage/homevisiting/sir02082011.pdf

defining the concepts (in this case, the "constructs" already listed in the guidance document) to be measured.[3]

The various steps in this process include:

- Defining a concept or construct
- Selecting or developing a measure or indicator (including type: input, output or outcome measure)
- Developing an operational definition for each measure (including what will constitute improvement)
- Selecting data source(s) and measurement tool(s) or instrument(s), if needed
- Developing a data collection plan
- Collecting the data
- Analyzing the data
- Acting on the findings as part of a ongoing quality improvement process

As mentioned above, the Supplemental Information Request (SIR) already identifies concepts or constructs (e.g., breastfeeding, utilization of emergency department services) that need to be measured within each of the benchmark areas. The guidance document, however, provides discretion to grantees to select or develop a specific indicator for each construct in order to capture a dimension of the construct that is most useful to individual grantees. For example, if maternal depression is the construct in question, one possible indicator among others for this construct might be the percent of clients screened (i.e., percent of newly enrolled mothers screened for depression utilizing a valid and reliable tool during one year).

There are various *types* of indicators available to capture each construct. The types of indicators may be based on the components of the logic model of the home visiting program, i.e., inputs, outputs and outcomes. In the example of maternal depression, availability of mental health providers for referral of at-risk clients would be an indicator or measure of input; the number or proportion of clients screened would be an output type of indicator, and the percent of clients who screen positive and are therefore at high risk for depression would be an outcome type of indicator. Grantees have discretion to choose the type of indicator for each construct that is most useful or appropriate for their programs.

The Value of Selecting Appropriate Indicators and Measure Tools

It is important not only to conceptualize *what* is being measured but also define *how* it will be measured. Once an indicator is selected, it needs to be operationally defined. An operational definition is "a description, in quantifiable terms, of what to measure and the specific steps to measure it consistently."[4] Some indicators may require a measurement tool specified in the operational definition (e.g., a depression scale utilized to screen program participants such as the Beck Inventory or the Edinburgh scale) and others may not. In either instance, it is important to specify a data collection plan including: the person(s) responsible for collecting the data, the collection frequency, data sources, and method utilized.

[3] R Lloyd: "Indicator Selection and Development." Chapter 3 of *Quality Health Care, A Guide to Developing and Using Indicators.* 2004.
[4] Ibid.

The use of appropriate measurement tools is necessary to be able to collect the best evidence of program results; accordingly, the selection of a methodologically sound and appropriate measurement is essential. The remainder of this paper focuses on the selection and development of indicators and the need to specify an operational definition as well as the considerations in planning for the collection of data and selecting a methodologically sound (i.e., valid and reliable) measurement instrument when the indicator selected requires one.

Step 1: Selecting Indicators and Defining Measurable Improvement

The definition or development of indicators can vary across individual programs given the varying nature of program services and preferred goals and objectives across program models. Carefully considering how each indicator is defined and aligned with a desired objective for your program will allow you to (a) meet federal accountability requirements for the construct and benchmark area; and, (b) utilize changes of the indicator over time for your own purpose to continually improve your program's processes and outcomes.

Indicator Type

There are various ways of categorizing indicators. Two common types of indicators are process and outcome measures. Process measures capture program services and activities, programmatic policies and procedures implemented. Process measures collect "output" data that are used to describe who receives program services, what they received, and the frequency and intensity of services provided. For instance, process measures may be used to assess changes in:

- Number or percent of women screened for maternal depressive symptoms
- Number of clients who were referred for substance abuse treatment

Outcome measures are developed to measure change in child, family, and system-level outcomes. While process measures are useful in tracking program implementation, outcome measures are useful at examining change at the client level. Outcome data is often collected to assess improvements or changes in participant knowledge, attitudes, skills, or behaviors. Outcome measures may be used to assess changes in:

- Depressive symptoms among caregivers
- Positive parent-child interactions
- Knowledge of child development
- Caregiver use of alcohol or illegal substances
- Recurrence of child maltreatment within the family
- Length of time families receive public assistance

Considering the type of indicator used to capture a given construct is a key step to meet both accountability and continuous quality improvement goals. Depending on how the indicator is defined and the measurable objective (improvement) set, some constructs would be best captured with process measures while other constructs with outcome measures. In some instances, even goals predicated on improving an input type of measure could be appropriate (e.g., increasing the number of partners providing mental health and substance abuse services for referral in the community). The selection of type of indicator may vary across grantees for a given construct depending on how each grantee sets the measurement objective (improvement).

Step 2: Developing an Operational Definition of the Indicator

After selecting an input, output or outcome-related indicator that is the most appropriate for the related construct for your program, it is crucial to develop an operational definition for the indicator. An operational definition is "a description, in quantifiable terms, of what to measure and the specific steps needed to measure it consistently".[5] A thorough operational definition

- Gives communicable meaning to the concept
- Is clear and unambiguous
- Specifies the measurement method, procedures, and measurement tool (when needed)
- Provides decision-making criteria when necessary
- Enables consistency in data collection [6]

Grantees should define the specific components of the indicator (e.g., numerator and denominator if it is a percentage or a rate). The definition should also include any measurement tool needed to capture the data. Also, indicators may or may not have targets or goals to track improvement.

The SMART (Specific, Measurable, Attainable, Relevant and Timely) goal system is a useful resource in developing an operational definition of an indicator that includes the desired objective (or improvement). [7] The following is an example of an indicator written by a program serving pregnant women utilizing the SMART goal method.

Construct: Prenatal Care

Objective: Increase the rate of pregnant women served by the program who receive at least one prenatal care visit in the first trimester from year 1 baseline to the 3-year benchmark reporting period

Possible data sources: Interview, self-report surveys, administrative records

---

This objective for the indicator is specific and measureable. The definition specifies the service population of focus (pregnant women served by the program) and identifies what is being measured – the timing of the start of prenatal care among pregnant women in the program. The indicator includes a definition of improvement with a well-defined time frame. The program will measure pregnant women in the program at baseline, and again at year 3 to look for favorable change (improvement) in the rate.
Once an indicator and what constitutes measurable improvement have been operationally defined, the next step is to plan the data collection effort and determine the data sources that are the best fit for your program.

---

[5] *Ibid.*
[6] I*bid*
[7] For more information on setting SMART goals and objectives see O'Neil, J. and Conzemius, A. (2005). *The Power of SMART Goals: Using Goals to Improve Student Learning*. Bloomington, IN: Solution Tree Press.

Step 3: Data Collection Planning: Selecting Data Sources

Data collection requires planning. Unless thought is given to identifying the persons responsible for collecting the data, the frequency of collection and its cost, the data sources, and the methods (e.g., sampling) to be utilized, the validity of the findings may be challenged or their usefulness questioned at a later time.

Data Source

According to the federal Supplemental Information Request (SIR), benchmark area-related data should include all families participating in the program. To measure changes among families in the program, the collection of individual-level data is therefore required. Individual-level data refer to demographic, service utilization, and outcome information that is collected about each individual participating in the program. Individual-level data allow programs to look at findings as they relate to specific demographic and other characteristics of program participants (e.g., families enrolling prenatally or postpartum), and to examine patterns across participants who have been served by the program. Although data collection at the individual-level is needed, MIECHV grantees will report to the Federal government data only at an aggregate level (i.e., summary data for all families served by the program).

To select the data source that is most appropriate for your program, consider the following issues:

- What source(s) is/are likely to provide the most accurate information?
- What source(s) is/are the least costly or time consuming?
- Will collecting information from a particular source pose an excessive burden on that person?
- What are the steps required to access the data?
- How often will the data be collected? [8]

To adequately measure all of your program constructs, it is likely that you will need to implement multiple data collection methods. Data sources can include self-report interviews and surveys, direct observations, or administrative records. Several factors need to be considered before deciding on a data collection method.

For example, if a construct will be measured with data collected through interviews, a protocol to adequately train interviewers should be implemented to (a) ensure relative consistency in the administration of the protocol and (b) avoid potential sources of measurement error, such as deviations from the interview protocol. However, since interviews may provide higher item-by-item response rates than other data sources, the payoff may justify the added burden. If using self-report surveys, bear in mind that cost varies across assessments. It is also important to develop a plan to address response rates for those measures that may have lower response rates than other data sources. Data collected without some participants may bias the aggregate data reported later.

[8] For more information of factors to consider when selecting a data source, see The Program Manager's Guide to Evaluation. (2010). Washington DC: Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved March 28, 2011, from: http://www.acf.hhs.gov/programs/opre/other_resrch/pm_guide_eval/reports/pmguide/pmguide_toc.html.

Data collected through direct observation requires thorough training of staff to ensure the accuracy and consistency of the data collected. Finally, data collection from administrative records can be very time consuming but it can provide very valuable data. Collecting administrative data also removes the burden of data collection off front-line staff. However, there are often many unforeseen barriers to accessing administrative data maintained by other local or state human service entities.[9]

These barriers should be anticipated to the extent possible and addressed early in the process to ensure the accuracy and quality of the data available. For example, administrative data may appear to measure an indicator you are interested in, but key individuals who know the administrative data well may inform you that the administrative data does not, in fact, match the operational definition set in the context of the home visiting program.

Determining the best data source to build the indicator (and capture an important dimension of the construct of interest) will depend on both program and contextual factors. For example, to determine the number of emergency room visits of a client in the last six months, one program may choose to collect the data from all clients through self-report interviews while another may opt to pull the data from administrative medical records. While self-report data may be easier to collect, it may also be less reliable than medical records data. In contrast, medical records data may be more accurate, but it may involve more staff time to collect and may necessitate engaging other parties and overcoming legal or bureaucratic constraints than self-report data. [10]

Step 4: Reviewing and Selecting a Measurement Tool

Once measurable improvement has been defined and an indicator has been selected, certain indicators require the further selection of a measurement tool or instrument. The next step, therefore, is to review and select a measurement tool that provides the best fit for the indicator and the program. Consider these factors when selecting a measurement tool:

- Is the tool standardized (normed)?
- What are the training requirements for administration, scoring and interpretation?
- Is it reliable?
- Is it valid?
- Is it sensitive to assessing change, and specific to what you are measuring?
- What are costs involved?
- Can the data be used by staff for continuous quality improvement?
- Is it appropriate for the population of focus (e.g., norms, linguistic equivalence, etc.)?

[9] For more information on accessibility issues in human service database or similar information management system, see James Bell Associates. (2009, September). Evaluation brief: Common evaluation myths and misconceptions. Arlington, VA: Author.

[10] The guidance to states strongly encourages grantees to use consistent indicators within a benchmark area across home visiting models if more than one home visiting model is implemented within a state (i.e., a single indicator for each construct across implemented models). In addition, the guidance encourages the data collection across benchmark areas be coordinated and aligned with other relevant state or local data collection efforts.

A brief overview of each element is provided below.

Standardization of Measurement Tools

An important consideration to make when examining a possible measurement tool is to determine if the instrument is standardized. A standardized instrument is designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent across administrations and participants. Simply, it means the same test is given in the same manner to all subjects all the time. It also means there is an established protocol for scoring and interpretation of the results.

While standardized measurement tools are generally preferred over non-standardized ones because they are easily administered on a wide scale and scoring and interpretation procedures are readily available, they are often inadequate for describing subgroups. They may also be culturally or linguistically biased or less appropriate for a specific population. If you are using a standardized test, it is important to consider the norm group by which program participants will be compared to ensure proper interpretation of findings. A norm group refers to the group of people who have already taken the assessment during the development and testing of the measure; the performance of your participants will be compared to the performance of the norm group which, hopefully, will be an appropriate group for comparison purposes. Since participant scores will be compared to the norm group scores, it is important to establish similarities and identify any meaningful distinctions between program participants and the norm group.

Another caution when using standardized instruments– even though an assessment is standardized, the results may not be meaningful if the measure is administered, scored, or interpreted inappropriately. One example is administering a timed test under untimed conditions. A second example is if staff are not trained adequately, resulting in inconsistent administration to participants, which can bias the data collection process.

Training Requirements

Another important factor to consider in the process is the level of training required to administer the measures. Whether using a measurement tool with highly specific training requirements or one without specified training requirements, training is still necessary and relevant before administering a measure. A detailed protocol should be established regarding how to administer an assessment and how to handle the data once it is collected. Additionally, it is helpful to have periodic booster trainings to “refresh” staff on the data collection protocol. This ensures the consistency of administration across staff as well as over time. Some training protocols include “spot-checking” of staff performance during administration to ensure that protocols are being administered as specified by all staff.

Reliability of Measurement Tools

Another factor to consider when selecting measurement tools is the reliability of the measure. Reliability estimates the consistency and stability of your measurement, or more

simply the degree to which an instrument measures the same way each time it is used under the same conditions with the same subjects. For example, a reliable developmental screening tool administered to the same participants in the same manner within 24 hours should yield nearly identical results (based on the assumption that child development does not change significantly in a 24 hour period). Thus, reliable measures imply the repeatability of findings. There are several methods to assess reliability: test-retest, split-half, inter-rater, and internal consistency.

Test-Retest Method – This is a relatively simple method of estimating the reliability of a measurement tool, in which the same instrument is administered to the same individuals at different points in time. The higher the positive correlation is between the two scores, the higher the estimate of the reliability of the measure. Calculated reliability scores based on the correlation between the two instrument administrations range between zero and 1.00, with scores closer to 1.00 indicating less error variance and higher degrees of reliability. In general, there are four cut-off points for reliability, which include excellent reliability (0.90 and above), high reliability (0.70-0.90), moderate reliability (0.50-0.70) and low reliability (0.50 and below).[11] The advantage of the test-retest method is that it is relatively simple; however, there are several limitations.

The greatest concern with the method is the lack of repeatability of some measures with the same subjects over time. For example, changes in respondents' attitudes or behaviors may occur over time, affecting the correlation of the measurement scores at different points in time. In this case, lack of correlation between the measurement scores may largely be attributable to changes in the respondents themselves rather than reflect the measure's reliability. In general, the longer the time interval between measurements, the more likely it is that respondents' attitudes and behaviors will have changed, leading to differences in test responses.

Another concern with the test-retest method involves how respondents react to the measuring tool itself. In this case, being exposed to the measure leads to a change within the respondents that will affect the respondent's responses to the measure at the second administration. Thus, differences between the obtained scores collected at the two separate time intervals are the result of the respondent's sensitization to the measure itself rather than an indication of the reliability of the measure.

Split-Half Method – While the test-retest method requires two administrations to the same respondents at different time intervals, the split-half method requires one administration to a single group of respondents. In this method, the responses are divided in half and the correlation between the two halves is used to estimate the measure's reliability. Since the obtained correlation estimates the reliability of both halves of the assessment rather than the entire assessment, a statistical correction must be made to estimate the reliability of the entire assessment.[12] Just as with the test-retest method, calculated reliability scores for the split-half method range between zero and 1.00, with scores closer to 1.00 signifying strength

[11] For more information see Hinton, P.R., Brownlow, C., McMurray, I. and Cozens, B. (2004). *SPSS explained.* East Sussex, England: Routledge.

[12] For more information on statistical corrections for the split-haves method, see Carmines, E.G. and Zeller, R.A. (1979). *Reliability and Validity Assessment: Sage University Paper Series on Quantitative Application in the Social Sciences, 07-017.* Newbury Park, CA: Sage.

of the relationship and higher degrees of reliability. In general, reliability scores of .70 or higher are considered acceptable, and scores of .90 and higher are considered excellent.

Concerns related to the split-half method of estimating reliability are related to varying methods of grouping items into halves, since the method used to determine the split will lead to slight variations in scores compared to alternative methods. Methods used to determine the split include randomly assigning items into two groups, separating even- and odd-numbered items, and dividing the first and second half of the assessment into groups. Therefore, the split-half method may yield slightly different reliability scores depending on the method used to determine the split; this is true even though the same measure is administered to the same individuals during a single administration session.

Internal Consistency Method – The internal consistency method compares the different items within the same instrument to ensure they are measuring the same dimension of a construct. The method involves a single administration of the same measurement tool to the same group of respondents. Cronbach's alpha is then applied; it generates a statistical correlation coefficient to estimate the degree of internal consistency of a measurement tool. Internal consistency scores can range from zero to one, with 1.00 indicating 100% correlation. among the items in the instrument. Generally, a Cronbach's alpha score of .70 or greater indicates acceptable reliability, and 0.90 or higher indicates excellent reliability.

Inter-Rater Reliability or Inter-Observer Reliability Method – Unlike the previous measures of reliability which examine the consistency of the measurement tool, inter-rater reliability measures the consistency of the administration of the instrument. Using this method, the same measurement tool is administered to the same respondents at different time intervals using different data administrators. The correlation between the observed scores of the two data administrators will estimate the reliability or consistency between the data administrators.

There are several implementation methods that can be used to help ensure the reliability between data administrators. These include: establishing a thorough and clearly-defined data collection protocol, training and re-training staff periodically, and supervised shadowing of data administrators. These methods, while not directly measuring reliability, help ensure consistency in the administration of measures.

Validity of Measurement Tools

A valid measure is one that measures the concept it was intended to measure. Validity, then, refers to the accuracy of your measurement. For example, if a construct is operationally defined as improved children's social-emotional development, but the measure selected focuses strictly on physical development, the measure is not a valid measure of social-emotional development. While the measure may be reliable, it is not valid because the tool is ***not*** measuring the construct it was intended to measure. There are several types of validity.

Criterion-Related Validity - Criterion-related validity measures the extent to which a measurement tool predicts an outcome based on some external criterion or indicators of a construct. There are two types of criterion-related validity – concurrent (the extent to which a measure predicts present behavior) and predictive (the extent to which a measure predicts future behavior). Predictive validity is determined by the degree to which a measurement

tool accurately predicts future performance or behavior in relation to the construct being measured. Another way to measure criterion-related validity is to determine how well the measurement tool's results correlate with those of other established measurement tools that assess the same construct. For example, if you were interested in using an instrument that assesses maternal depression, you could administer the instrument along with other established scales of maternal depression. This would allow the observed results of the measurement tool of interest to then be correlated with other established instruments for this same construct. Correlation coefficient scores can range from zero (0.00) to one (1.00), with higher scores indicative of a stronger relationship between the two instruments for the same construct. A validity score of .60 or above is considered moderately high.[13]

Construct Validity – Construct validity refers to the extent to which a measurement tool accurately captures and samples the domain of behaviors associated with a theoretical conceptualization of a construct. A measure that demonstrates construct validity is one that accurately represents the given construct and provides a comprehensive and accurate sample of the domain of behaviors encompassing that construct. Construct validity is based on the logical relationship among variables. A measure has construct validity if it demonstrates an association between the observed scores obtained from the measure and the prediction of a theoretical trait. Correlation coefficient scores can range from zero (0.00) to one, (1.00), with higher scores being suggestive of a stronger relationship between measures. A validity score of .70 or above is considered moderately high.

Content Validity - The extent to which the measurement tool reflects the range of the possible skills or behaviors that make up the construct being assessed determines the content validity of the measure. Elements that make up the defined construct are used as the criteria by which the measurement tool is compared. There is no statistical test that can be conducted to determine content validity. Rather, the content validity of an instrument is determined by a thorough review of the instrument by experts on the topic/content area that the tool is purported to measure. The purpose of this thorough review is to ensure that all relevant areas of the construct are being addressed by the items included within the measurement tool. For example, if using a cognitive development measurement tool to assess childhood cognitive development, the instrument should be comprised of items that adequately sample childhood cognitive development. The items should cover the full range of intended domains of cognitive development. There should be a sufficient number of items included in the instrument to ensure that obtained scores credibly represent cognitive development in those domains. Otherwise, a lack of a sufficient number of items poses a potential threat to content validity since aspects of the full domains comprising the construct may be underrepresented.

Programs should determine whether a measurement tool has been demonstrated to be valid for its particular purpose. It is recommended that programs choose those measures that are demonstrated to have as many types of validity as possible. Face validity, or the idea that a measurement tool looks like it would measure what it intends to measure, is not sufficient to establish validity for measurement tools. The other types of validity discussed above should be discussed in the selection of a measurement tool.

[13] For more information see Kaplan, R.M. and Saccuzzo, D.P. (2001). *Psychological Testing: Principle, Applications and Issues (5th Edition).* Belmont, CA: Wadsworth.

Reliability and Validity

Reliability and validity are related concepts, both of which have implications to the quality or accuracy of a measuring tool. A measuring tool cannot be valid unless it is also reliable. By contrast, a measuring tool can be reliable independently of its validity: an instrument can be reliable without being valid. Additional evidence demonstrating that the instrument is indeed measuring what it is purported to measure must be available to verify the validity of an instrument, etc. Reliability, therefore, is a necessary but not sufficient condition for validity. The following figure provides an analogy of the relationship between reliability and validity.

Figure 1. Reliability and Validity[14]

The first illustration presents an example of a shooter that is reliable but not valid. The shooter is consistently hitting the same area of the target, but always missing the bull's eye. To be a valid measure, you have to hit the mark, so to speak, or measure what it is you are trying to measure. If an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently.

The fourth illustration is analogous to a measure that is both valid and reliable. The arrows are consistently hitting the bull's eye. Not only does it accurately measure what you what, but the results are repeatable.

A final note about reliability and validity: it is necessary to understand the conditions in which an assessment has been found to be reliable and valid. For example, if a measurement tool has been deemed both reliable and valid in English only, translation of the measure into another language affects the reliability and validity estimates of that measure. Programs assessing clients whose primary language is not English must use measures demonstrated to be reliable and valid in the language in which they will be administered. Likewise, if a multiple-scale measure has been demonstrated to be reliable and valid in its entirety but not by each separate scale, administering the scales separately changes the psychometric properties obtained for the entire measure.

Sensitivity and Specificity

[14] For more information see Shuttleworth & Martyn (2008). Validity and Reliability. Retrieved 2/18/11 from Experiment Resources: http://www.experiment-resources.com/validity-and-reliability.html.

The sensitivity of a measure refers to the degree to which an instrument correctly identifies those individuals who have a specific condition. A sensitive measure has the ability to detect differences between groups. For example, if you're testing for developmental delays, a highly sensitive assessment will correctly identify developmental delays among participants. The level of sensitivity of a measure is determined by the proportion of people who identify with a given condition. For example, if 20 children in a given sample have a developmental delay, and the measure accurately identifies 18 of the children as having a developmental delay (18/20=.90), then the measure is 90% accurate at detecting developmental delays in children. A highly sensitive measure of developmental delays is not likely to let a child with a developmental delay fall through the cracks. On the other hand, highly sensitive measures often have a tendency to inaccurately identify children without a developmental delay as having one.

The specificity of a measure refers to the degree to which an instrument correctly identifies those individuals who do not have a specific condition. Specificity refers to the ability of an instrument to correctly "screen out" those individuals who do not have a specific condition. The level of specificity of a measure is determined by the proportion of people who correctly screen out for a given condition. For example, if 100 children in a given sample are ***not*** developmentally delayed and the measure screens out 85 of them (85/100=.85), then the measure is 85% accurate at screening out children for developmental delays. While a highly specific measure will accurately screen out those children without a developmental delay, they often have a tendency to inaccurately screen out children who actually may have a developmental delay.

The higher the sensitivity and specificity of a given measure, the greater the accuracy of that measure. Ideally, measures should have high levels of both sensitivity and specificity. In reality, however, there is often a tradeoff between sensitivity and specificity. There is no general consensus about what constitutes acceptable levels of sensitivity and specificity. What is considered "acceptable" will depend on issues such as the intent of the measure, the level of risk for the given condition, the prevalence of the condition in the group being tested, and available alternate methods of assessment.

Costs Associated with Measurement Tools

Another important consideration when deciding on a measurement tool is to consider the costs associated with the measure. There are both financial and time-related costs. First, what are the financial costs associated with using a measure? It is important to assess the cost to purchase the assessments, train the staff and, possibly, to hire appropriate staff to collect the data prior to selecting measures.

It is also important to anticipate costs associated with time prior to making decisions on measurement tools. Important questions to consider include estimating how much time staff will spend implementing and using this measurement tool; determining whether this tool can replace another tool; identifying the value of the tool for the program's continuous quality improvement efforts; and anticipating costs to determine if the benefits provided by the use of a particular measure justify the added costs.

Utility of Scores for Staff

Ideally, collecting data to assess benchmarks should be integrated in the program's continuous quality improvement efforts. Data collection can be seen as an opportunity not only to meet funding requirements, but also to review ongoing results to assess the quality and effectiveness of program implementation. Focused and critical assessment of program implementation can provide programs with a unique opportunity to build upon identified strengths and focus on areas requiring improvement. Ideally, data collected to assess benchmarks will be reviewed by the program at all levels to assess progress and identify areas for growth.

Appropriateness to Population of Focus

The final consideration when selecting measures is to consider the appropriateness of the measurement tool for your population of focus. Is the measure culturally and linguistically appropriate for the population(s) you are working with; is it age or developmentally appropriate; are there any language barriers or literacy barriers that can affect the administration and/or interpretation of the measure relative to the population(s) you are serving? For example, if using a norm-referenced instrument, the norming samples ought to include the ages, culture, and language of the children in your program. Often, assessment manuals and technical documents provide descriptions of the populations that were used for norming. Information provided about the normed groups (keeping in mind how long ago the norming was done) should be compared to the population of focus.

Often, measures are used with cultural groups or specific subpopulations which do not include normative data and psychometric properties about those cultural groups or subpopulations. For example, a language development screening assessment may have established its reliability and validity with native English-speaking populations; if it is translated into Spanish for use with a predominantly Spanish-speaking population, however, proper psychometric analyses would need to be performed to ensure that the reliability and validity (as well as the semantic and linguistic equivalence) of the translation was appropriate for use with this cultural group as well. Furthermore, adaptation to the original assessment, such as translating in Spanish, should be done with care to ensure that cultural biases are addressed and the translated version is relevant and understandable to the subpopulation. Language translations should be done by professionals skilled in working with the subpopulation.

Of all the considerations previously discussed, determining whether a measure is appropriate to use with your population of focus can, arguably, be the most important factor to consider. A program can thoughtfully address all the required considerations by selecting a standardized measure that is reliable and valid, adequately training the staff, and implementing the prescribed data collection protocol consistently. If, however, the measure is not culturally or linguistically appropriate for the program participants served, then the measure is inadequate, regardless of the value of the measure with other populations. Thus, judiciously addressing the unique characteristics of the population of focus is crucial to the measurement selection process.

Conclusion

A wide range of factors should be considered when determining how the individual constructs or concepts within each benchmark area will be assessed. This multi-phase process involves selecting an indicator related to each construct (including type of indicator), developing an

operational definition incorporating a definition of improvement (with or without the setting of targets), and selecting measurement tools (when they are needed) based on key factors that affect the quality and appropriateness of the instrument selected. Planning the data collection process with attention to the person(s) responsible for collecting the data, the frequency and method utilized, and the selection of optimal data sources are also important steps.

Using appropriate indicators and measurement tools is necessary to be able to accurately and consistently collect evidence of successful program results. The importance of good measurement is brought home by the understanding that the findings resulting from the these indicators will inform subsequent action to continuously improve your program. Accordingly, careful consideration is needed at all steps throughout the process of indicator and measurement tool selection.

For more information about assessing constructs and selecting appropriate measures, please contact a DOHVE[15] TA team member at:

Susan Zaid, MA
TA Liaison
James Bell Associates
1001 19th Street, North, Suite 1500
Arlington, Virginia 22209
703-528-3230 or 800-546-3230
www.jbassoc.com

Virginia Knox, PhD
Project Director
MDRC
16 East 34th Street
New York, New York 10016
212-340-8678
www.mdrc.org

[15] The purpose of the Design Options for Home Visiting Evaluation (DOHVE) is to provide research and evaluation support for the Maternal, Infant and Early Childhood Home Visiting (MIECHV) Program. The project is funded by the Administration for Children and Families in collaboration with the Health Resources and Services Administration.