The Vocabulary of Program Evaluation

Program evaluation, like every field, has its own vocabulary. This vocabulary ensures that everyone working on the evaluation shares a common understanding when discussing program evaluation and its concepts and procedures. Keeping the language of program evaluation in mind as you plan for and conduct the evaluation will help keep you on track toward producing a quality product. Some of the terms may be familiar to you, while others may be new. Use this handout as a reference throughout the program evaluation process.

Averages: Mean, Median, Mode, and Outliers

An ability to calculate *averages* is important for working with and understanding the significance of sets of data, such as that illustrated in figure A.1.

$$\begin{array}{r} 1{,}000 \\ +1{,}000 \\ +1{,}000 \\ +1{,}000 \\ +1{,}000 \\ +2{,}000 \\ +3{,}000 \\ +3{,}000 \\ +3{,}000 \\ +3{,}000 \\ \underline{+100{,}000{,}000} \\ 100{,}019{,}000 \end{array}$$

$$100{,}019{,}000 \div 11 = 9{,}092{,}636.36$$

Figure A.1: Calculating the mean of a sample data set.

Mean

The *mean* is what most people think of when they say *average*. To calculate the mean, add all of the numbers in a set, and then divide by the number of figures in the set. In figure A.1, there are eleven numbers in the set.

Median

The *median* is that point above which and below which half of the group falls. In figure A.1, the median is 2,000, because half of the numbers are larger than 2,000, and half are smaller.

Mode

The *mode* is the number that occurs most frequently in a set. In figure A.1, the mode is 1,000.

Outliers

An *outlier* is an extremely low or high value in a set of numbers. In figure A.1, that number is 100,000,000. Including an outlier in your data set can skew the results. For example, the mean of the numbers in figure A.1 when the single 100,000,000 figure is included is 9,092,636.36. Eliminating that outlier, the calculation becomes 16,000 ÷ 10, which at 1,600 is much more representative of the range of numbers in the list. It is a good idea to determine how the group will handle outliers before beginning data collection (Bracey, 2003).

Charts and Graphs

An easy-to-use resource for understanding strengths, differences, and how to create various data displays is the Maryland Department of Education's School Improvement in Maryland website (n.d.) (http://mdk12.org/instruction/curriculum/mathematics/math_guidelines.html).

At this site you will find guidelines and examples in creating the following displays:

- Bar graphs
- Box-and-whisker plots
- Circle graphs
- Frequency tables
- Line graphs
- Scatter plots
- Line plots
- Pictographs
- Stem-and-leaf plots

Cognitive Demand

Cognitive demand (sometimes referred to as *depth of knowledge*) is the level of thinking required of a student in order to complete a given assignment. In order to provide students with the knowledge and skills necessary for them to meet a particular grade-level standard, instruction must be aimed at the level of cognitive demand specified in the standard. Similarly, in order for an assessment to result in accurate information about the progress of a student or group of students toward the standard, the cognitive demand of the assessment must match the cognitive demand of the content standard. Table A.1 provides an overview of different levels of cognitive demand as measured by Bloom's Revised Taxonomy.

Table A.1: Cognitive Demand as Measured by Bloom's Revised Taxonomy of Educational Objectives

Higher-Order Thinking Skills	
Creating	Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing
Evaluating	Making judgments based on criteria and standards through checking and critiquing
Analyzing	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing
Lower-Order Thinking Skills	
Applying	Carrying out or using a procedure through executing or implementing
Understanding	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining
Remembering	Retrieving, recognizing, and recalling relevant knowledge from long-term memory

Source: Anderson & Krathwohl, 2001, pp. 67–68

Fidelity of Implementation

Fidelity incorporates two concepts: "*adherence* to the intervention's core content components and *competent* execution using accomplished teaching practices" (Forgatch, Patterson, & DeGarmo, 2005, p. 3). This means that the program, practice, or strategy is fully carried out as designed by the creator or program provider. It also implies implementation by individuals experienced in the program's delivery.

Fidelity of implementation is essential to the success of a program because

> the way a program is implemented influences whether the program will be effective, or not effective. Poor implementation or lack of fidelity of implementation can, and often does, change or diminish impact. Also, once a program has been modified, no one quite knows how it will operate or what unexpected consequences it might produce. (California Department of Education, 2006, p. 1)

Formative and Summative Program Evaluation

Just as in student evaluation, program evaluation uses two kinds of assessments: *formative* and *summative*. One of the clearest descriptions of the difference between the two is Robert Stake's analogy: "When the cook tastes the soup, that's formative; when the guest tastes it, that's summative" (as cited in Frechtling, 2002, p. 8).

As the analogy illustrates, formative assessment is used to sample the program as it exists at that moment, when time still remains to modify it. Summative assessment is used to answer questions about the impact of the program. It cannot be used to change what has already occurred, but it can be used to modify future renditions of the program.

Generalizability

Generalizability is the concept behind representative sampling. It is the concept that conclusions based on the responses of a sample population can, in certain circumstances, be extended to the population as a whole. The larger the sample, the more generalizable the findings. When sample sizes are small, the less generalizable the findings. This has meaning in school-level program evaluation because sample sizes will range from large to small. For example, if the evaluation is looking at a math intervention program used in the primary grades, interviewing only one class out of three at each grade level will not give an adequate sample to be generalizable. A sample of one is simply not large enough to ensure that the findings are generalizable. However, if in a school of 300 students, 270 (90 percent) responded to a survey, it is likely that the results would be generalizable to the other students.

Goals

Goals are public statements that define how good *good enough* is. These statements must include measurements that describe how the school community will know when the goal has been reached and a time frame within which the goal will be met.

Inter-Rater Reliability

Inter-rater reliability refers to the consistency in scoring when more than one person is involved in that scoring. Inter-rater reliability is strong when different raters obtain the same or very close to the same scores when using an identical data collection instrument with the same individual in the same circumstances. This is a consideration when measures such as classroom or other observations are made by more than one observer. It means that all of the raters assign scores that are the same, or very close to the same, no matter who does the observation.

Objectives

Objectives are checkpoints or milestones on the pathway toward meeting the goal. Objectives must be measurable and have a *reach by* date so that the school community is able to gauge progress toward the goal.

Program Evaluation

Program evaluation is a systematic, organized process to provide evidence on the effectiveness and pros and cons of programs, practices, or strategies. It can be large scale—for example, evaluating all reading programs in a district or state—or small scale—such as evaluating the math program in a single school or even within a single grade level. While large-scale evaluations are most often carried out by specialists from outside the school system, small-scale evaluations are most frequently conducted by a school team led by or under the direction of the principal.

Research Questions

Research questions are the questions whose answers will address the purpose of the program evaluation. These questions must be specific and measurable or observable.

Quantitative and Qualitative Data

Data used in program evaluation fall into two broad categories:

1. **Quantitative**. These data focus on numbers and other types of information that can be measured, such as how many eighth-grade students met the National Assessment of Educational Progress standards in math.

2. **Qualitative**. These data explore that which can be observed but is not easily measurable, such as people's attitudes and beliefs.

To illustrate, consider figure A.2, which illustrates the differences between the quantitative and qualitative descriptions of a latte.

Quantitative Description of a Latte	Qualitative Description of a Latte
Serving temperature of 150 degrees	Enticing aroma
Ingredients: 1.25 cups of milk, 2 tablespoons of vanilla flavored syrup, and 1.5 fluid ounce of brewed espresso	Appears light tan and frothy
180 calories for 12 fluid ounce latte made with whole milk, 100 calories if made with skim milk	Comforting taste of steamed sweet milk and espresso
9 grams of fat made with whole milk, 0 grams of fat if made with 2% milk	Tastes even better when syrups are added to create different flavors
More than 300,000 sold daily	Goes well with chocolate chip cookies

Figure A.2: Comparison of quantitative and qualitative observations.

Reliability

Reliability is the extent to which data gathering would yield the same results if the evaluation were repeated by the same people under the same conditions. For the purposes of program evaluation, this is important because the most effective way to increase reliability is to ensure that the methods used for data collection are consistent. For example, suppose one program evaluator consistently runs out of time when conducting interviews and hurries through or even skips the last three questions, while another interviewer allocates an equal amount of time to each question. It is likely that this inconsistent data collection will skew the results obtained because the respondents had unequal times within which to respond to each question.

Subgroup Populations

One facet of program evaluation is to examine whether a particular program, practice, or strategy is effective with *all* children as a group, *and* with smaller groups of children who share common characteristics. The characteristics used to subdivide student populations include the following:

- Gender
- Grade level
- Participation in 504 plans
- Receipt of special education services
- Socioeconomic status

Validity

This means the data collection methodology will actually measure what it is supposed to measure. For example, let's assume that a program evaluation is measuring the cognitive demand of student assignments. If the data collection consisted of comparing the assignments to the levels of Bloom's Revised Taxonomy of Educational Objectives, the measure would have validity because each assignment's level of cognitive demand would be measured according to a taxonomy of higher- and lower-order thinking skills. If, however, the data collection consisted of comparing the assignments to the teacher's lesson plans, the measure would not have validity because the data collected would describe how well the assignments matched the lesson plan, but not the degree to which the assignments called for higher-order thinking skills.