GUIDANCE FOR PRE- and POST-TEST DESIGN

The simplest evaluation design is pre- and post-test, defined as a before & after assessment to measure whether the expected changes took place in the participants in a program. A standard test, survey, or questionnaire is applied before participation begins (***pre-test or baseline***), and re-applied after a set period, or at the end of the program (***post-test or endline***). Pre- and post-tests can be given in writing or orally.

The goal of this guidance is to help programs avoid some of the most common errors in use of pre- and post-evaluation. More detailed guidance is available in "Useful Resources" listed below.

For example, if the launch of an after-school program teaching business skills coincides with the addition of these skills to the local school math curriculum, it would be hard to know which results are due to the program and which to the school. The program could respond by analyzing where the program overlaps with the school curriculum, and focusing on program results in those areas that the school curriculum does not address.

The main weakness of pre- and post-test design is that it cannot detect other possible causes of positive or negative results among the participants. For this reason, if a new program is under consideration for expansion, other explanations for the results should be ruled out by collecting more information on each possible explanation, and then singling out the results due to the program.

Design Review

In cases where a new pre- and post-test is being written, or an existing test adapted, **pilot testing is strongly recommended.** A good method of piloting would be to convene a group of youth advisors -- from the same communities as youth who will be in the program -- to take the test, discuss it among themselves using this checklist, and suggest modifications. This is a perfect opportunity for youth participation and leadership.

<u>Tips for developing and reviewing pre and post questions</u>

1. **RELEVANCE OF CONTENT TO OBJECTIVES:** Does the content clearly address the objectives of the program?
    - If your program focuses on knowledge, then match content as nearly as possibly to the learning objectives.
    - If your program aims to change attitudes or norms, do the questions cover the range of attitudes that your program addresses?

2. **LENGTH** – The shorter the better, especially if there are open-ended questions.[1] Eliminate redundant questions. Pilot test to ensure taking the test would take no more than ½ hour.[2]

3. **EDUCATIONAL LEVEL** – Ensure that the reading or vocabulary level is at the right level for the youth participants. Determine whether literacy levels demand oral interviews. Feedback from youth on unfamiliar or ambiguous words or phrases is helpful.

> In English, the reading level can be checked in MS Word. Run "Spelling/Grammar"; and the last item under "Readability" that shows up is the Flesch-Kincaid Grade Level analysis. Other tests might be more reliable in other languages.

4. **CULTURAL OR LANGUAGE ADAPTATION**– When pilot test-takers cannot agree on the meaning of a question, adaptation is needed. Questions on sensitive issues such as reproductive and sexual health often must be adapted to the local youth culture. On the other hand, the same test/questionnaire applied to adults should not use the terms from youth culture, which adults may find offensive.

5. **AVOID OVERLY GENERAL OR AMBIGUOUS QUESTIONS:** Questions that are too general are subject to a variety of interpretations, giving inconsistent results.

    - For example, "Do you think girls and boys should be treated equally?" is not as clear as a specific question about the goals of the program such as "Do you think girls should be able to play [local sport] in public places?"

6. **AVOID LEADING OR BIASED QUESTIONS:**
    - A leading question may lead the respondent into a pre-determined answer that may not accurately reflect their opinion. For example: "How has your life changed as a result of the program?" should be changed into "Has your life changed in any way as a result of the program?" (yes, no), followed by multiple choices or open-ended response.
    - A biased question will lead the participant to give a socially acceptable response. For example, "Do you drink too much alcohol at parties?" should be changed to "Do you drink alcohol?" and then if yes, give a range of choices on frequency, amount, and setting.

7. **AVOID ASKING TWO QUESTIONS IN ONE**: For example, "How would you rate your financial knowledge and skills?" should be changed into two separate questions.

8. **MIX POSITIVE AND NEGATIVE STATEMENTS** when measuring attitudes or behavior through statements asking respondents to "agree" or "disagree". Randomly mix statements that reflect the attitudes promoted by the program versus those that are discouraged. For example, if a gender program's post-test only has statements favoring gender equality, respondents will detect easily the desired response.

9. **SAMPLING:** When programs serve large numbers of youth, often there are not enough staff or funds to apply the pre- and post-test to all of them. In that case, the evaluators generally decide on their sample size,[3] and use two methods to achieve a non-biased representative sample, that is, a smaller set of youth who are likely to reflect the characteristics of the larger group:

---

[1] Open-ended questions do not define a set of responses, so that participants come up with their own answers. An example would be: "What is your favorite activity?"

[2] See suggestions for eliminating questions on page 32 of Barkman under Useful Resources" below.

[3] https://www.surveymonkey.com/mp/sample-size-calculator/

- "Random selection" ensures that each youth in the program has an equal chance of being chosen. Computerized methods are commonly used.[4] One non-computerized method is like a lottery. For example, in a program involving 500 11th grade students, put their names on slips of paper into a bowl, and draw names one by one until the desired sample of 50 for the pre- and post-test is reached.
- "Systematic selection" is another way to avoid bias. Divide the total population of youth in the program by the sample size you have decided, then use the resulting number *n* to choose every nth student. Using the same example, from the complete list of the 500 students in grade 11, pick every 10th student on the list to be tested, and end up with the sample of 50.

Tips for Analysis & Reporting of Results

A sample pre- and post-analysis template in Excel is available from your EMpower program officer, if that would be helpful. The calculations for each student as well as for the whole group are in formulas both for the numerical change in score and for the percent change.

1. ***Use analysis of the pilot test to identify questions that need to be improved.*** Analysis of the pre- and post-tests might catch other needs for improvement. Common signals of needs for improvement are:
    - Many blank answers to specific questions
    - Inconsistent answers to related questions
    - Responses to open-ended questions that reflect lack of understanding
    - Failures to respond to the questions towards the end of the test (respondent fatigue due to length).
2. ***Use analysis of pre-test scores as a guide to curriculum:*** If more than 60% of students answer certain questions correctly for knowledge (or in the desired direction for attitudes), your program is unlikely to have a major effect on these items. Use the results to adjust your curriculum to focus more on the areas where most students scored low, and consider removing the questions where they scored high from the post-test.
3. ***Analysis of quantitative data:*** Most instruments yield quantitative data from close-ended[5] questions or ratings. The template EMpower has developed is designed for quantitative data analysis.
    - These data are generally analyzed to compare pre- and post-tests for frequencies, such as per cents and averages.
    - Statistical analyses are needed to look at changes over time, and the significance of differences between pre- and post-tests.
    - If you don't have access to statistical expertise, especially when the numbers of participants are greater than 50, it would be helpful to enlist a local researcher to analyze significance.[6]

[4] https://www.randomizer.org/tutorial/ and https://www.surveymonkey.com/blog/2012/06/08/random-sample-in-excel/
[5] "Close-ended" ended questions provide set choices; types include multiple choice, true-false questions, and questions using rating scales. See pages 20-25 of the Barkman guide on these options, referenced below.
[6] To identify which changes from baseline to endline are probably due to the program, and not to chance. Most significance tests aim for the probability of results due to chance being 5% or less.

4. ***Reporting quantitative data:*** In a final report, always present results in numbers as well as percentages in a table, with a final column comparing the increases or decreases between the pre- and post-test. Never report only per cents, without a reference to the number of respondents. See example below.

SAMPLE TABLE

	Pre-Score	Post-Score	Difference	% Change
Female Average (N=15)	28.47	36.27	7.80	27%
Male Average (N=15)	24.27	33.27	9.00	37%
TOTAL AVERAGE	26.37	34.77	8.40	32%

   - ***In the discussion of results***, for specific items in your survey, point out: 1) findings that represent significant results, and what these strengths mean for your program; 2) any differences (if applicable) between male and female participants or other sub-groups among your participants, and how you plan to address any disparities in results by group; and 3) where results were not as good as expected, discuss how your program plans to address these areas.

5. ***Reporting qualitative data:*** Open-ended questions yield qualitative data, which may be analyzed by theme, or through applying scores that can reported on quantitatively.[7]
   - Identify common themes: Report the themes from a significant number of youth in their answers. For example, in a sample of 50 students, any themes or ideas put forward by more than 10 are worth reporting. An illustrative quote or two is useful to illustrate these common themes or ideas (without identification of the persons who provided the quotes).
6. ***Pulling it all together:*** Compare the qualitative and quantitative data findings: Does one set of data support or raise questions about the analysis of the other set? Taken together, what do these two sets of data mean for your program?

Resources

Susan J. Barkman, "A Field Guide to Designing Quantitative Instruments to Measure Program Impact." An excellent basic publication from Purdue University.
http://www.northskynonprofitnetwork.org/sites/default/files/documents/Field%20Guide%20to%20Developing%20Quantiative%20Instruments.pdf

E. Jane Davidson's work is an excellent guide to evaluation in general, and is known for explaining how to use rubrics to analyze qualitative information in evaluating outcomes.

a. Davidson, E. Jane, *Evaluative Reasoning* http://www.unicef-irc.org/publications/pdf/brief_4_evaluativereasoning_eng.pdf
b. Davidson, E. Jane mini-books: http://realevaluation.com/read/minibooks/
   - *Making the Important Measurable, Not the Measurable Important*

[7] See E. Jane Davidson in Useful Resources below.

- *Actionable Evaluation Basics*

I-Tech, *Guidelines for Pre- and Post-Testing*, has a short and useful discussion of both design and analysis of pre- and post-tests, focused in their examples on knowledge questions. Available on request from program officers.

a. I-Tech has a toolkit with many other resources for evaluation of training programs. See http://www.go2itech.org/HTML/TT06/toolkit/evaluation/index.html The forms tab has many tools which could be used pre- and post-test, including written as well as observation checklists.