School Leadership Program (SLP)

Summary of Comments from Grantees on Instructions for Annual Performance Report (APR)

GPRA Measure 2.1

Comment on scoring:
Our measure for Objective 2, Performance Measure 2.1, the McRel Balanced Leadership Profile, yields a profile score, not a unitary score, which is the type of data requested for the performance measure. Also, on this measure, some scales assess "progress" by an increase in scores, while others assess it by a decrease in scores. Our measure, like the Val-Ed from Vanderbilt, is a 360 measure (requires others to rate the principals in addition to the principal's response). For these reasons, data from our measure do not fit neatly into the "effect size" spreadsheet provided.

We mentioned above that the construct of leadership is multidimensional, which means that a unitary score will not capture what is being measured. New trends in leadership assessment include outside ratings of the leader to capture multiple perspectives as a balance to the self-report of the leader. This makes for multiple scores on multiple score scales. All of the information can be handled statistically, but not with the methods proposed in the Annual Performance report (ED 524b) form.

Response to comment on scoring:
No changes were made to the APR form or instructions based on this comment because for this performance measure, it should be sufficient to report the principal score only from the instrument used by the commenter. The commenter compares their instrument to the Val-Ed, one of the assessments suggested by the Department. It should be noted that Val-Ed does provide an overall effectiveness score for each respondent.

Comment on reliability:
None of the suggested measures have reported reliability scores with the exception of inter-item correlations. Reliability for internal consistency is dependent upon having a "lot" of items within a category. When a profile is used, internal consistency can drop, as the number of items for each profile will be fewer than the total number for the test. The alphas reported for Val-Ed reflect the fact that there are 72 items in the "sample." Our chosen instrument, the McRel Balanced Leadership Profile, has 21 scales on a (roughly) 76-item assessment that would yield too few items per scale to show a high internal consistency. Since our assessment (and Val-Ed) is a 360 assessment, the appropriate reliability would be "interrater reliability," a special case of generalizability. This reliability isn't reported for Val-Ed. It could be calculated for our instrument.

Response to comment on reliability:
The APR instructions will be revised to direct grantees to contact program staff with questions on how to report and explain the data collected from grantee project-developed instruments. No additional changes to the APR form or instructions were made based on this comment because gathering and reporting additional evidence would place considerable burden on grantees in terms of expertise and financial resources.

Comment on validity:
The issue of validity is central in our critique of the pre-post measure. What is the construct of leadership and how do we know we're measuring it with our instrument? Both Val-Ed and McRel see leadership as multidimensional qualities/competencies, best assessed by several people to give multiple perspectives on the "leader." By definition, a multidimensional construct cannot be characterized by a single score. Secondly, validity is wedded to purpose. A test is not valid in and of itself. The only "validity" a test possesses is from how accurate the decisions made from test scores prove to be. Thus, a measure's validity must be determined by how accurate the results are for subsequent decisions. In leadership programs, the test use needs to be taken into account when assessing validity. Are we trying to predict some kind of "on the job" success? Are we trying to predict which skills, knowledge, and beliefs are changing? Are we trying to predict whether a high score on the test leads to some observable changes in a school's climate, achievement, collaboration, teacher satisfaction? These "validities" are all different and require different kinds of validity evidence. Empirical evidence for any of these test purposes is lacking for the suggested assessments. Our chosen assessment, the McRel balanced Leadership 360, has predictive validity obtained through meta-analysis for changes in student achievement, a primary goal of the project.

First, the proposed measures do not meet the criteria for reporting reliability and validity. Second, the Val-Ed, the strongest of the proposed measures, is not conceptually amenable to a unitary score. It contains multiple scales. Should field trials for Val-Ed (commencing in 2009) find that all the scales are highly correlated enough to yield a score, it could either mean there were too few items for each scale to yield high inter-item correlations and/or they are measuring some construct underlying all the scales which we might call "leadership". This construct could be defined differently for each different kind of leadership test. Thus the validity evidence linking scores to other predictive outcomes would be essential to determine what "kind" of leadership we are assessing.

Given the ancient state of leadership assessment, we would suggest that the 524b form ask for the kinds of evidence, internal consistency, interrater reliability, predictive validity, etc. that each project has obtained from its chosen leadership measure and how that evidence supports accurate decisions about leaders observed during the life of the project. The cross-project meetings could share their information so that we can build on the very slight research on leadership assessment.

Response to comment on validity:
The suggested instruments on the list were reviewed to ensure that they were developed based on a theory and/or research and have at least some evidence as to the validity of the scores. The APR instructions were revised to specify components related to evidence of content validity for grantees that are developing their own assessment or administering an alternate assessment not on the suggested list. Grantees will be asked to provide the following: purpose of the instrument; description of what the scores mean; rationale for instrument items (e.g., theory items are based on); and review process (e.g., how items were reviewed for bias and alignment with theory). No additional changes to the APR form or instructions based on this comment because gathering and reporting additional evidence would place considerable burden on grantees in terms of expertise and financial resources. The GPRA measure is intended to be a gross measure of program performance and is not intended to measure impact or to be used to make decisions about individual participants.

Comment on effect size calculation:
The effect size calculation needs revision. First, if there were a unitary score for these measures, the small sample sizes and the lack of variability of the sample in leadership programs (you need normally distributed data and a rather hefty sample size to calculate standard deviation information) would undermine this procedure. Second, the reason for using effect size is to get at "practical significance". There are several methods for doing this, confidence intervals, for example. And when using a leadership profile, the proper statistic would take into account "shifts" in the profile rather than shifts in a unitary score. I suspect there are ways to do these measures. However, sample sizes and instrument reliability will probably make this a difficult task.

Response to comment on effect size calculation:
No changes to the APR form or instructions were made based on this comment. We will analyze first year data to determine if the concerns listed above are relevant.

Comment on pre-/post- testing:
In order to ensure that we capture the growth of second year principals who do not return for a third year of coaching, we will do this. However, we feel it will be duplicative for the 3rd year principals who were coached in their second year. They will then be taking a post-test in June and another pre-test in September, with little to no SLP coaching support over the summer.

Response to comment on pre-/post-testing:
The APR instructions were revised to allow grantees that provide services to a cohort over multiple years and that conduct a pre-test and post-test the first project year to use the year one post-test as the pre-test for the second year and, as appropriate for subsequent years. Also, grantees are asked to define program completer for their SLP project to clarify when final data will be available on those participants considered to have completed the program

GPRA Measure 2.2

Comment on tracking retention for program participants:
We are aware that we should track retention for two years after the participants complete coaching. However, we feel that this should only be assessed for principals who are enrolled in coaching for both their second and third year as principal.

Response to comment on tracking retention for program participants:
The APR instructions were revised to ask grantees to define program completer for their SLP project and allow grantees to limit tracking to participants considered to have completed the program.