Progress Monitoring Technical Review Committee

Frequently Asked Questions

1. What does the TRC expect vendors to submit for GOM 2: Reliability of the slope?

Reliability of the slope measures the ratio of true score variance to total variance. An explanation of this concept can be found in Raudenbush and Bryk (2002). Page 50 has an equation that describes how reliability can be computed (equation 3.59). Reliability of the slope is different than standard error of the slope in that its value is not dependent on sample size.

For example, vendors might do one or more of the following. *Please note that the method for calculating the reliability of slope should match the procedures users will actually use when calculating slope.* Also note that these examples are not exhaustive; they are simply suggestions from the TRC.

(1) Use the software program HLM to obtain the estimate of reliability for the slope. HLM will produce estimates for reliability of the slope automatically. There is a student version of this program available free of charge at http://www.ssicentral.com/. Fitting a model that allows for random intercepts and random slopes will more than likely be sufficient for obtaining this reliability estimate.

(2) Use another program, such as SPSS or SAS to run a mixed modeling command to produce estimates for reliability of the slope. This is more cumbersome since these programs do not produce reliabilities estimates and additional calculations must be made. The generic formula for reliability is (true score variance)/(total variance). To compute reliability of a slope using SPSS or SAS, one needs to obtain estimates for each.

To obtain estimates of the true score variance of the slope estimate, one needs to run a procedure for mixed models. In SPSS, it's called Mixed, and in SAS, it is Proc Mixed. Both programs require that your data be in univariate format (where your dependent variables such as Oral Reading Fluency would be represented in one column or variable, and the rows represent the number of occasions an assessment was made on each person. So if a person has four assessments, for example, there would be 4 rows of data for that person). Another variable indicating time at which the assessment occurred is also needed. The metric of this time variable does not matter. It could be in days/weeks/months from the first assessment, for example. So if one has a dataset that has three variables (ID, ORF, and TIME) laid out in a univariate format, one could run the following code in SPSS

```
MIXED orf with time
 /Print=solution testcov
/method=ml
```

/FIXED = time

/random =intercept time |subject(id) covtype(un).

Or in SAS,

```
proc mixed data=temp covtest;
  class id;
  model orf=mtime;
  random intercept mtime /subject=id type=un;
run;
```

From this, you can obtain the estimated true score variance of the slope. In SPSS, it is the estimate labeled UN(2,2) in the box "Estimates of Covariance Parameters" and in SAS it is also labeled UN(2,2) in the section labeled "Covariance Parameter Estimates".

This is the estimate of the true score variability of the slope, and will serve as the numerator of your reliability estimate.

The denominator for your reliability is the estimate of total variance of the slope, and can be obtained in a number of ways. What you need is to obtain the OLS slope estimate for each person, then compute the variance of these estimates. If you are familiar with SAS, these can be obtained directly in the Proc Mixed procedure, or one can simply run a regression for each person (predicting their performance using time), then taking the unstandardized regression weights for each person for time and find the variance of those weights. Then use this as the denominator of your reliability estimate.

(3) Produce a split-half reliability estimate by calculating a slope of improvement through each student's odd scores; calculating a slope of improvement through each student's even scores; and producing a correlation between the two slopes.

Note that for all of these options, the TRC asks that vendors to provide information on

(a) the number of data points per student (average and range);

(b) the number of months spanned by the data collection per student (average and range) included in the analyses.

Citation:
Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods (Second Edition). Thousand Oaks, CA: Sage Publications.

2. What does the TRC expect vendors to submit for GOM 4: Predictive validity for the slope of improvement?

To provide information on the predictive validity for the slope of improvement, vendors should

correlate the value of the slope with an achievement outcome. The achievement outcome needs to be (a) external to the progress-monitoring measure/system and (b) either concurrent with the last data point used for the slope or better would be a measure delayed in time from the last data point used for slope. Vendors also need to specify what the measure is and when, in relation to the last progress-monitoring score, it was collected.

The TRC asks that vendors provide information on

(a) the number of data points per student (average and range);

(b) the number of months spanned by the data collection per student (average and range).

If the outcome measure used is not external to the progress monitoring measure/system, then the vendor must describe what provisions have been taken to address the limitations of this method, such as possible method variance or overlap of item samples.

3. The protocol asks if vendors have disaggregated reliability and validity data *in their user's manuals*. If these data are available but not reported in the technical manual, will you still evaluate our disaggregated data?

The TRC encourages vendors to make public any information they have about the reliability and validity of their tool for students in different racial-ethnic groups; hence the request that vendors indicate whether or not disaggregated data are included in their published manual. In its dissemination efforts, the Center will inform consumers of the availability of these disaggregated data. If you have disaggregated data available but have not published them in your manual, please do include them as part of your submission. We will ask you to consider making these data available to consumers.

4. What does the TRC expect to see for GOM 5: Alternate Forms?

For a full bubble rating on GOM5, the vendor must demonstrate that there are at least 20 alternate forms, AND that mean performance on alternate forms is comparable; in other words, the forms are reasonably equivalent. It is not sufficient to indicate equivalence across a subset of the forms; evidence must be included that demonstrates comparability across all 20 (or more) forms. It is also not sufficient to simply describe the construction process for these forms. Actual empirical data must be submitted to support form equivalence.

Evidence submitted for this GOM can take various forms. Some examples include:

(a) Descriptive statistics and test of comparability for form differences. For example, a mean and range of coefficients (e.g., means) for all possible forms, and a one-way ANOVA comparing the mean scores of the forms.

(b) A model-based estimate of form effects or reliability, such as from generalizability theory or multilevel models. Variance components or intraclass correlation from multilevel models (e.g., persons nested within forms) can be efficient to show the relative size of mean differences due to many forms (and other design effects, if present).

5. What does the TRC expect vendors to submit for GOM 6: Rates of Improvement?

Rates of improvement specify the slopes of improvement or average weekly increases, based on a line of best fit through the student's scores. The TRC is looking for growth standards that are based on normative or criterion-based evidence.

An acceptable response to this question would be a table specifying the growth standards, such as the examples below:

Example 1.

Grade	**ROI**
1	1.36
2	1.17
3	1.03
4	0.81
5	0.89

Example 2.

	Designating Risk		**Determining Response**	
Sample size	Level	5-8 Week Slope	Projected End-Year Benchmark	Slope of Improvement
Grade 1: 202	<5	<1.75	50	2.0
Grade 2: 324	<15	<1.00	75	1.50
Grade 3: 309	<50	<0.75	100	0.75
Grade 4: 326	<70	<0.50	125	0.50
Grade 5: 179	<80	<0.40	130	0.40
Grade 6: 247	<90	<0.35	150	0.35
Grade 7: 136	<90	<0.30	150	0.30

The developer must justify the method for establishing these standards for average weekly growth. Strong evidence on this GOM would also provide contextual information on the instructional conditions under which the standards were established.

6. For GOM 6, what sample should be used to estimate weekly growth?

The protocol asks vendors to describe the characteristics of their normative sample demographically. In addition, the TRC requests that vendors indicate in their response whether estimates of weekly growth are based on (a) an academically representative sample, or (b) a sample that has received intervention, or (c) a low-achieving sample that has not received intervention.

When estimating weekly growth, the TRC suggests considering the following:

(1) The TRC requires 20 alternate forms, so weekly measures, collected across at least 20 weeks, are encouraged (biannual or 3-time "benchmark" measures are not sufficient).

(2) The shape of the growth trajectory.

(3) Attention to reliability of the slope estimate (see the TRC's guide for slope reliability, in FAQ 1).

7. What does the TRC expect vendors to submit for GOM 8: Sensitive to Student Improvement?

Sensitivity is the extent to which a measure reveals improvement over time, when improvement actually occurs. For this GOM, the TRC is looking for evidence from empirical studies that evaluate the extent to which score improvement on the measure is related to score improvement on other measures. For example, an acceptable response to this question could be that an empirical study found that slopes on the progress monitoring tool are significantly different than zero; the slopes are significantly different for learning disabled vs. low-achieving vs. average-achieving vs. high-achieving students; and that the slope are greater when effective practices are in place.

Strong evidence on this GOM would also provide contextual information on the instructional conditions under which the measure's sensitivity was evaluated. Additionally, although not required, evidence demonstrating that the tool is sensitive to *validated interventions* would be considered especially strong.

8. How is sensitivity to student improvement determined in a computer adaptive test (CAT)?

Sensitivity is evaluated by the TRC for a CAT, based on the evidence submitted regarding how small the measurement error is and evidence regarding growth. With respect to measurement error, CATs have the ability to minimize error based on the response patterns of individual students--a set stopping rule is used so only as many items are needed to get a set level of precision regarding a student's score. In this respect, CATs are efficient, and if the stopping rule is good, precision is also very high. This is evaluated from the technical evidence provided to the TRC. The evidence regarding growth is evaluated in the same way as for other measures, such as the extent to which change in CAT scores relate to change in other scores, or group-randomized comparison studies. So while a CAT has the potential to minimize error in scores, those scores get evaluated for sensitivity to growth in the same way as for any other instrument.

From the description of CAT above, the IRT-based scoring algorithm makes a test more efficient: only enough items are used to reach a set level of precision (lack of error). Usually a CAT guarantees to use the minimum number of items specifically targeted to the student's ability, thereby reducing boredom, fatigue, and frustration. CAT algorithms can be changed, but only with permission of the software programmers. Thus, there is the potential for a CAT to be far more precise than any paper and pencil test, but this depends on the item selection algorithms used (e.g. what level of error is tolerated vs. maximum test length). Generally speaking, CAT is to be preferred because of its precision, efficiency, as well as virtually unlimited number of test forms available. Whether this translates into sensitivity to growth is the separate but related empirical question the TRC tries to evaluate for a given test.

9. What does the TRC expect vendors to submit for GOMs 9 and 10: Decision Rules for Changing Instruction and Decision Rules for Increasing Goals?

The purpose of GOMs 9 and 10 is to identify and evaluate the evidence on which decision rules for changing instruction and increasing goals are based. Strong evidence for these GOMs would include an empirical study that compares a treatment group to a control and evaluates if student outcomes increase when decision rules are in place.

10. What does the TRC expect vendors to submit for GOMs 11 and 12: Improved Student Achievement and Improved Teacher Planning?

The purpose of GOMs 11 and 12 is to identify and evaluate evidence that use of the tool results in positive outcomes, including improved student achievement and improved teacher planning. Strong evidence for these GOMs would include an empirical study that compares a treatment group to a control and evaluates if student or teacher-related outcomes improve when the progress monitoring tool is being used.

11. What is the difference between GOMs 11 and 12?

The difference between these two GOMs is in the outcome variable. GOM 11 refers to the impact of using the tool on student achievement and GOM 12 refers to the impact of using the tool on teacher planning. For GOM 12, therefore, the study would need to include a dependent variable that measures teacher planning. Examples include, but are not limited to: (a) alignment between teacher's short-term instructional targets and student needs (as revealed in objective data sources); (b) number of program adjustments; (c) number of increases in student goals; and (d) expert ratings of the quality of teacher plans.