Academic Progress Monitoring Frequently Asked Questions (FAQ)

1. **How does the TRC consider evidence for tools that can be used at multiple grade levels?**
   Submissions must report data separately for each grade level that are targeted by the progress monitoring instrument. Evidence will be rated and reported on the chart separately for each grade. In cases where data are not available for one or more grades that fall within the grade span targeted by the tool, the TRC will give a rating of "—" to indicate "data were not available."

2. **What is the difference in requirements between the "performance level standards" section of the chart and the "growth standards" section of the chart?**
   For data reported on the first tab of the chart ("performance level standards"), vendors will be required to report analyses conducted on the general population of students (i.e., a sample that is representative of students across all performance levels). For data reported on the second tab ("growth standards"), vendors will be required to report analyses conducted on a population of students in need of intensive intervention. Convincing evidence that children were in need of intensive intervention may include one or more of the following: all students below the 30th percentile on local or national norm or sample mean below 25th percentile on local or national test; students have an IEP with reading goals or math goals that are consistent with the tool; or students are non-responsive to Tier 2 instruction.

3. **What does the TRC consider sufficient with respect to sample size?**
   For each of the technical standards, rather than specify a concrete minimum sample size, the TRC has established a lower bound for an estimate, and requests that the vendor provide a confidence interval (CI) around the estimate. If a sample is small but evidence shows that the estimate remains above this lower bound, it will be considered acceptable. This lower bound varies by standard and is stated in the rating rubric. Note that for model-based psychometric evidence, providing Test Information Function (TIF) or Standard Error (SE) plots to judge the relative precision of an estimate are acceptable in place of providing CIs.

4. **What does the TRC expect vendors to submit for reliability of the performance level score, and what factors are considered when rating the quality of this evidence?**
   The TRC considers reliability analyses that are rigorous, and that are appropriate given the type and purpose of the tool.

   For progress monitoring tools which use total scores or fluency-based measures, the TRC recommends reporting model-based indices of item quality. These can include McDonald's omega (Dunn, Baguley, & Brunsden, 2013; McDonald, 1999) for categorical SEM or factor models, Item Response Theory estimates of item quality based on item information functions (Samejima, 1994), or reliability of the score for fluency-based measures. For IRT-based

models, vendors should consider reporting marginal reliability as well as an ability-conditional estimate (e.g., report reliability estimates for students with differing levels of ability) so that the strength of IRT reporting can be fully leveraged in reporting (Green, Bock, Humphreys, Linn, & Reckase, 1984). Note that for marginal reliabilities, coefficients may not differ much from Cronbach's alpha and can therefore be interpreted using the same guidelines.

If model-based approaches are not used, it is expected that strong evidence for at least two other forms (see list of examples below) of appropriately justified reliability are provided to receive a full bubble. Regardless of the type of reliability reported, given that intended uses for tools can vary, it is incumbent on the vendor to provide supporting justification of choice of emphasis for reliability evidence.

Examples of Forms of Reliability:

- Alternate form:
    - For multiple forms, evidence can be provided to indicate that the alternate forms yield consistent scores across probes within a given set (e.g., using median score of multiple probes) and across time period.
- Internal consistency (alpha):
    - Ad hoc methods for item-based measures include internal consistency methods such as alpha. Alpha is the mean of all possible split halves (Cronbach, 1951). However, alpha is not an index of test homogeneity or quality per se (Schmitt, 1996; Sijtsma, 2009).
- Test-retest:
    - Test-retest data should be provided with a minimal time period of 1 week (no more than two)
- Inter-rater:
    - The TRC strongly recommends that inter-rater reliability be reported for tests which are subjective and require human judgment (e.g., open-ended questions as opposed to simple choice selection or computer recorded responses that would not require inter-rater reliability. The analyses should acknowledge that raters can differ not only in consistency, but also in level. Possible analyses include multilevel models of ratings within judges and students, generalizability theory, and invariance testing in SEM.

*Note that the TRC ***does not accept*** split-half reliability metrics. Split half reliability is problematic given that these methods can be arbitrary and potentially artefactual. Additionally, the TRC ***does not recommend*** that vendors submit test-retest as a reliability metric. Test-retest is problematic given that high and low retest reliability may not always signal a reliable assessment, but instead reflect student growth patterns (e.g., high test-retest can mean that students aren't changing over time, or maintaining the same rank order, and low test-retest can mean that students are meaningfully changing over time and changing differently).

5. What does the TRC expect vendors to submit for validity of the performance level score, and what factors are considered when rating the quality of this evidence?

The TRC expects vendors to offer a set of validity analyses that offer theoretical and empirical justification for the relationship between its tool and a related criterion measure. In other words, the vendor needs to specify the expected relationship between the tool and a criterion, and then use an appropriate empirical analysis to test this relationship. The TRC discourages vendors from providing a large list of validity coefficients correlating with multiple criterion measures, and instead recommends a few analyses that have a theoretical basis about a relationship between the tool and a small set of appropriate criterion measures.

Types of validity may include: evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and/or evidence based on consequences of testing. The vendor may include evidence of convergent and discriminant validity. However, regardless of the type of validity reported, the vendor must include a justification demonstrating how these data taken together demonstrate expected relationships between the measure and relevant external criterion variables. If appropriate, the vendor should take into account the fact that analyses against more proximal outcomes might be expected to show higher correlations than analyses against distal measures and offer explanations of why this is the case.

In the TRC's view, validity evidence based on ***internal*** criterion measures (i.e., measures that come from the same "family" or suite of tools) does not provide adequate support for validity. The TRC expects evidence based on criterion measures that are ***external*** to the progress monitoring system. The TRC encourages vendors to carefully select external criterion measures, and recommends choosing other, similar measures that are on the tools chart. If it is necessary to use internal measures, the vendor must describe provisions that have been taken to address limitations such as possible method variance or overlap of item samples.

6. How does the TRC consider evidence that is disaggregated for demographic subgroups (e.g., English learners, students with disabilities, students from different racial/ethnic groups)?

The TRC encourages vendors to include data disaggregated for demographic subgroups. Any submission that includes disaggregated data will be noted on the chart with a "d" superscript, and users can access the detailed information by clicking on the cell. Note that disaggregated data will not be rated, but instead just made available to users. An advanced search function for the chart will also enable users to quickly locate tools that have data disaggregated for the subgroups in which they are interested.

7. What kind of evidence does the TRC expect to see for Bias Analysis?

With respect to bias, the greatest threat to validity is construct-irrelevant variance (Messick, 1989, 1995) which may produce higher or lower scores for examinees for reasons other than the primary skill or trait that is being tested. The issue of bias, or lack thereof, constitutes an argument for validity (Kane, 1992). Arguments for the valid use of a test depend on clear definitions of the construct, appropriate methods of administration, and empirical evidence of the outcome and consequences.

In general, comparisons of group means are not sufficient for demonstrating bias or the lack thereof because the properties of the items are conflated with the properties of the persons (Embretson, 1996; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Measurement models of latent traits (e.g., item response theory, confirmatory factor analysis, or structural equation models for categorical data) are better suited to provide rigorous examinations of item versus person properties. Speeded tests present additional complications, but those complications do not remove the need to understand issues of test fairness or bias.

The overarching statistical framework for issues of bias is that we have a structural factor model of how a trait predicts item responses (McDonald, 2000) and this model is tested for equality across two groups (Joreskog, 1979; Vandenberg & Lance, 2000). Most analyses of group differences can be seen as simplifications or restrictions on this general model. The TRC will consider any of the four methods below as acceptable evidence for bias analysis:

- Multiple-group confirmatory factor models for categorical item response (Meredith & Teresi, 2006). Categorical CFA allows the testing of equal item parameters across groups via a series of restrictions (e.g., from freely estimated to fully equated) to isolate group differences of persons from item bias.
- Explanatory group models such as multiple-indicators, multiple-causes (MIMIC; Muthen, 1988; Woods, 2009) or explanatory IRT with group predictors (De Boeck & Wilson, 2004; Van den Noortgate, De Boeck, & Meulders, 2003).
    - MIMIC models attempt to test the equivalence of item parameters, conditional on background characteristics or group membership (analogous to an ANCOVA, but for a factor model). Most forms of a MIMIC model represent a restriction of a multiple group CFA.
    - Explanatory IRT uses a multilevel regression framework to evaluate the predictive value of item and person characteristics. A series of models with increasing (or decreasing) restrictions can be fit to test conditional equivalence (or non-significance) of item or person difference parameters.
- Differential Item Functioning from Item Response Theory (DIF in IRT). There are several approaches to evaluating DIF across groups (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Zumbo, 2007), many of which are exploratory methods to uncover the possibility of group differences at the item level. Vendors might also consider referencing Meade's taxonomy of standardized effect sizes for DIF that allow you for interpretation of the practical impact of DIF (Meade, 2010).

8. What does the TRC expect to see for 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 from 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. The student version of this program can be obtained free of charge from the Scientific Software International website, which is 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.

9. What does the TRC expect to see for validity of the slope?

To provide information on the predictive validity for the slope of improvement, vendors should correlate the value of the slope with an external 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. For details on how internal/external measures are defined, refer to Question 5. 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.

10. What does the TRC expect vendors to submit for alternate forms, and what factors are considered when rating the quality of this evidence?

For a full bubble rating on for alternate forms, vendors 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 the alternate forms standard 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).

These methods may not be suitable for computer-adaptive tests (CATs), as such assessments do not rely on fixed forms. Relevant analyses for CATs might include:

- Providing data on how frequently item characteristics are evaluated for stability/drift; this may include evidence from previous studies evaluating stability and equivalence.
- A description of how items are administered to reduce frustration, fatigue, and boredom.
- An examination of item coverage and construct validity; e.g., describing how many items of each kind are presented to children or explaining the stopping criterion.
- Data on how many questions are in the item bank and what procedures/rules are in place to reduce under/over-exposure of items.

11. What does the TRC expect vendors to submit for decision rules for setting and revising goals and for changing instruction, and what factors are considered when rating the quality of this evidence?

The purpose of the decision rules for setting/revising goals and the decision rules for changing instruction standards is to identify and evaluate the evidence on which decision rules for changing instruction and increasing goals are based. Therefore, the TRC expects to see evidence that the tool can accurately detect small changes in performance during the time period that the tool specifies is necessary for users to make decisions. Strong evidence for these standards may include:

- Analyses of data establishing rates of improvement and sensitivity to improvement, that are based on a sample of students in need of intensive intervention and from whom progress monitoring data have been collected at least weekly over the period of time specified in the tool's decision rules, or
- An empirical study that compares a treatment group to a control and evaluates whether student outcomes increase when decision rules are in place.

12. Can I submit tools that can be used as screening tools for review by the progress monitoring TRC?

Yes, if the tool can also be used for progress monitoring purposes. Specifically, the tool must be able to reliably measure growth in an overall academic domain over a period of 20 weeks or more. For tools which measure narrower skills that can be mastered in shorter periods of time and that predict competency in broader academic domains (e.g., letter naming fluency, oral counting), the TRC recommends submitting for review by the screening TRC, to be consistent with the way in which users are most likely to use the tool.

References

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models : a generalized linear and nonlinear approach*. New York: Springer.

Cronbach, L. J. (1951). Coefficient alpha and and the internal structure of tests. *Psychometrika, 16*, 297-334.

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. doi:10.1111/bjop.12046

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Jöreskog, K. G. (1979). Simultaneous factor analysis in several populations. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models* (pp. 189-206). Cambridge, MA: Abt Books.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McDonald, R. P. (2000). A basis for multidimensional IRT. *Applied Psychological Measurement, 24*(2), 99-114.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11, Suppl 3), S69-S77. doi:10.1097/01.mlr.0000245438.73837.89

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229-244.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369-386. doi:10.3102/10769986028004369

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering where It Has Been, where It Is Now, and where It Is Going. *Language Assessment Quarterly, 4*(2), 223-233.