Behavior Progress Monitoring Frequently Asked Questions (FAQ)

1. **How does the TRC consider evidence for tools that can be used across multiple grade spans and/or has forms for different rates (e.g., teacher, parent, student)?**
   Submissions must report data separately for each span of grade levels that is targeted by the progress monitoring instrument, in accordance with developer guidelines about target grade spans or ranges (e.g., K-1 or K-3). Data must also be reported separately by informant, if appropriate for the tool (e.g., teacher, parent, student). Evidence will be rated and reported on the chart separately for each possible combination of grade span and informant (e.g., K-1 teacher, K-1 parent). In cases where data are not available for one or more grades that fall within the grade span targeted by the tool, or one of the available informant forms, the TRC will give a rating of "—" to indicate "data not available."

2. **What is the difference in requirements for the "foundational psychometric standards" section of the chart and the "progress monition for intensive intervention" section of the chart?**
   For data reported on the first tab of the chart ("foundational psychometric 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 ("progress monitoring with intensive population"), 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: students have ED label; students are placed in an alternative school/classroom; students have demonstrated non-response to moderately intensive intervention (e.g., Tier 2); or students have demonstrated severe problem behaviors (e.g., Tier 3), according to an evidence-based tool (e.g., systematic screening tool or direct observation).

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 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.

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, 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). 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, split-half):
    - Ad hoc methods for item-based measures include internal consistency methods such as alpha and split half. Split half methods are arbitrary and potentially artefactual. 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:
    - Tests which require human judgment (as opposed to simple choice selection or computer recorded responses) should report evaluation of 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 recommend that vendors submit certain common reliability metrics—specifically split half and test-retest. Split half reliability is problematic given that these methods can be arbitrary and potentially artefactual. 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.

It is important to note that to support validity, the TRC prefers and strongly encourages criterion measures that are ***external*** to the progress monitoring system. Criterion measures that come from the same "family" or suite of tools are not considered to be external to the system. The TRC encourages vendors to select 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 mean by sensitivity to behavior change, what kinds of evidence should vendors submit to demonstrate this, and what factors are considered when rating the quality of this evidence?

Sensitivity-to-change refers to the extent to which a measure can detect incremental changes in behavior within a short period-of-time. This is particularly important within problem-solving frameworks in which progress-monitoring data informs determinations of a student's responsiveness to intervention. Sensitivity-to-change represents the association between session-to-session changes in student behavior and the degree to which the measure accurately reflects these variations. Documenting an instrument's sensitivity to change requires consideration of the technical feature of the instrument's scores with particular focus on level and trend. When considering methods for documenting sensitivity to change, vendors must provide evidence (a) that behavioral change occurred, (b) the amount of change that occurred, and (c) the reliability of the change using either statistical or visual methods (Chafouleas, Sanetti, Kilgus, & Maggin, 2012; Maggin & Bruhn, 2017).

The TRC considers sensitivity-to-change as a unique concept from response-to-intervention though it is acknowledged that the burgeoning nature of the construct and related methods requires some flexibility for documentation. As such, the TRC will currently accept evidence of sensitivity-to-change based on individual responses to intervention as long as computation of the metrics are from idiographically collected data. Under the current guidelines, researchers have several methods at their disposal for documenting sensitivity to change. These methods can include either (1) single measure methods or (2) comparative methods. Descriptions of these broad categories, as well as the specific methods that fall within each, are provided in the following sections.

Single measure methods are used to document a particular measure's sensitivity to change. Each of these methods expresses the nature and/or extent of change that a measure of interest has captured. This change is NOT evaluated relative to any other measure or outcome, but rather to individual responding. Two specific single measure methods are described below:

- **Change metrics.** Change metrics are metrics that express change in a variable over time (Gresham, 2005; Olive & Smith, 2005). Gresham et al. (2010) recommended several metrics to document change sensitivity in progress monitoring instruments, including absolute change, percent of non-overlapping data (PND), percentage change, computation of effect size measures, and the reliability change index. Other statistics may be included in this group as well, including alternative nonoverlap statistics (Parker, Vannest, & Davis, 2011) and regression-based techniques among others (e.g., Pustejovsky et al., 2016 Computation of these metrics are collected idiographically and compare a student's response to different conditions. Typically, these conditions include a baseline and intervention phase, though conditions might also refer to natural modifications to the environment as long as there is careful documentation. Sensitivity to change is demonstrated through these metrics if they document observable change in student responding on the target variables between the conditions. Vendors are encouraged to select multiple metrics to document sensitivity because each index has a unique set of assumptions and provides evidence for different properties of the data.
- **Dynamic models.** Whereas the aforementioned change metrics provide a descriptive approach to documenting sensitivity to change, a class of statistical models can examine individual variabilities using longitudinal data (e.g., Wang, Zhou, & Zhang, 2016). Underutilized in the social sciences, dynamic modeling can assist vendors in documenting an instrument's sensitivity for an individual by providing time-dependent variation within single individuals (Hamaker et al., 2005). The use of dynamic modeling for evaluating an instrument's sensitivity requires collection of many data points for individual participants and therefore many vendors might not be able to use this approach. It is presented here as an option given its appropriateness for the task. Several dynamic modeling approaches are available including the traditional p-technique and more recent developments including dynamic factor analysis (Nesselroade & Molenaar, 2004) and dynamic Rasch modeling (Verhelst & Glas, 1993). Vendors using dynamic modeling to document sensitivity-to-change must describe the model and provide a rationale for its use.

Comparative methods are used to examine the extent to which the change documented via a measure of interest is similar to the change documented via some criterion measure. Whereas the

threshold for single measure methods in evaluating sensitivity to change is the documentation of *some* change, the threshold for comparative methods is documenting change that is similar to that of an alternative measure. Because comparative methods set a higher threshold for sensitivity to change, they are considered a more stringent form of sensitivity to change evidence.

- **Visual analysis.** Miller et al. (2017) provided an example for documenting sensitivity-to-change through visual analysis. This method requires concurrent idiographic collection and graphing of the measure of interest with another measure. In the Miller et al. (2017) example, the authors compared data collected with the Direct Behavior Rating Single Item Scale (DBR-SIS; i.e., the measure of interest) to data collected with systematic direct observation (SDO; i.e., the criterion). The resulting graphs provide evidence of incremental variability across sessions and allows visual analysts to determine if the level, trend, and variability across sessions is consistent between the instruments. Sensitivity-to-change is supported when the instruments represent similar patterns in the data.
- **Correlational analysis.** Correlational analyses can be combined with the change metric approach (see above) in evaluating the extent to which change documented through one measure is correlated with the change documented through another measure. Chafouleas et al. (2012) provide an example of such an approach. Within this study, two absolute change scores were calculated for each of the 20 student participants, expressing the degree of change in student behavior from baseline to intervention phases. The first of these absolute change scores represented change in DBR-SIS scores, whereas the second corresponded to change in SDO scores. Spearman's rho (ρ) coefficients were then calculated to examine the extent to which these two sets of absolute change scores were correlated with each other.
- Although less commonly used, multi-level modeling also affords a method by which to compare multiple methods in terms of documented change. Specifically, multivariate growth models could be used to examine the correlation between both (a) measure intercepts, permitting examination of the association between baseline starting points or intervention termination points (depending on variable centering), and (b) measure slopes, permitting evaluation of the association between increases or decreases in a variable over time.

The TRC acknowledges that there is currently not one accepted framework for documenting sensitivity to change and that the selection of methods will require vendors to consider issues related to the instruments construction, scoring rubric, and purpose. Vendors are provided leeway to select the methods most appropriate for their instrument though justification for the methods are required. Please note that TRC members might request additional clarification or metrics if the methods used inconsistent or unclear.

9. What does the TRC expect vendors to submit for data to support intervention change and for intervention choice, and what factors are considered when rating the quality of this evidence?

The purpose of the data to support intervention change 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 if student outcomes increase when decision rules are in place.

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

Yes. However, the evidence submitted must demonstrate its adequacy for progress monitoring. For example, there must be sufficient data points to demonstrate sensitivity to small behavioral changes in short periods of time, and reliability data must be appropriate for the intended use of the tool for progress monitoring.

References

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

Chafouleas, S. M., Sanetti, L. M., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using direct behavior rating single-item scales. *Exceptional Children, 78*, 491-505.

Cronbach, L. J. (1951). Coefficient alpha 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.

Gresham, F. M. (2005). Response to intervention: An alternative means of identifying students as emotionally disturbed. *Education and Treatment of Children, 28*, 328–344.

Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System-Teacher Form. *School Psychology Review, 39*, 364.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate behavioral research, 40*, 207-233.

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.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*, 301-321.

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.

Miller, F. G., Riley-Tillman, T. C., Chafouleas, S. M., & Schardt, A. A. (2017). Direct Behavior Rating instrumentation: Evaluating the impact of scale formats. *Assessment for Effective Intervention, 42*, 119-126.

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.

Nesselroade, J. R., & Molenaar, P. C. (2004). Applying dynamic factor analysis in behavioral and social science research. *The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks, CA: Sage*, 335-44.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322.

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.

Verhelst, N. D., & Glas, C. A. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395-415

Wang, M., Zhou, L., & Zhang, Z. (2016). Dynamic modeling. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 241-266.

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.