National Center on
INTENSIVE INTERVENTION
at the American Institutes for Research®

AIR®
Advancing Evidence.
Improving Lives.

IDEAs that Work
U.S. Office of Special
Education Programs

Behavior Progress Monitoring Frequently Asked Questions

1. How does the Technical Review Committee (TRC) consider evidence for tools across multiple grade spans and/or have forms for different raters (e.g., teacher, parent, student)?

Submissions must report data separately for each span of grade levels targeted by the screening instrument, according to developer guidelines about target grade spans or ranges (e.g., K–1, K–3). Data also must be reported separately by informant (e.g., teacher, parent, student), if appropriate for the tool. Evidence will be rated and reported on the Tools Chart separately for each potential combination of grade span and informant (e.g., K–1 teacher, K–1 parent). When 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 "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 must report analyses conducted on the general population of students (i.e., a sample representative of students across all performance levels). For data reported on the second tab ("Growth Standards"), vendors must report analyses conducted on a population of students in need of behavioral intervention. Convincing evidence that children require behavioral intervention may include the following: students have an emotional disturbance label; students are in an alternative school/classroom; students demonstrated nonresponse to moderately intensive intervention (e.g., Tier 2); or students 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 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 rigorous reliability analyses that are appropriate given the type and purpose of the tool.

Examples of the types of reliability the TRC expects to see submitted include the following:

- **Alternate Form:** For multiple forms, evidence must indicate that the alternate forms yield consistent scores across probes within a given set (e.g., using the median score of multiple probes) and across time periods.
- **Internal Consistency:** Ad hoc methods for item-based measures include internal consistency methods, such as alpha and split half. Split half[1] 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[2] data include a minimum time period of 1 week (and no more than 2 weeks).
- **Interrater:** Tests that require human judgment (e.g., open-ended questions) versus simple choice selection or computer-recorded responses must report evaluation of interrater reliability. The analyses should acknowledge that raters can differ in not only consistency but also level. Possible analyses include multilevel models of ratings within judges and students, generalizability theory, and invariance testing in structural equation modeling.

Vendors also may submit model-based approaches to reliability. With model-based approaches, strong evidence from one analysis with at least two sources of variance (e.g., time, rater) is acceptable to receive a full bubble. For screening tools that use total scores, the TRC recommends reporting model-based indices of item quality, such as McDonald's omega (Dunn et al., 2013; McDonald, 1999) for categorical structural equation modeling or factor models and item response theory (IRT) 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) to fully leverage the strength of IRT reporting (Green et al., 1984). For marginal reliabilities, coefficients may not differ much from Cronbach's alpha and, therefore, can be interpreted using the same guidelines. In evaluating sources of variance, a model-based approach might be founded on generalizability theory, in which researchers examine the influence of various screening-

[1] The TRC does not recommend that vendors submit certain common reliability metrics—specifically split half and test-retest. Split half reliability is problematic because these methods can be arbitrary and potentially artefactual.

[2] Test-retest is problematic because high and low retest reliability may not always indicate the assessment's reliability but instead reflect student growth patterns (e.g., high test-retest can mean that students are not changing across time, or maintaining the same rank order, and low test-retest can mean that students are meaningfully changing across time and changing differently).

related facets (e.g., time, rater, screener forms) on the generalizability and dependability of the scores.

Regardless of the type of reliability reported, because intended uses for tools can vary, the vendor must provide supporting justification of choice of emphasis for reliability evidence.

4. 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 with theoretical and empirical justification for the relationship between its tool and a related criterion measure. In other words, the vendor must 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 an extensive list of validity coefficients correlating with multiple criterion measures; instead, the TRC recommends a few analyses with 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 (a) response processes, (b) internal structure, (c) relations to other variables, and/or (d) the 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 that demonstrates how these data, taken together, show expected relationships between the measure and relevant external criterion variables. If appropriate, the vendor should consider the fact that analyses against more proximal outcomes might 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 requires criterion measures that are **external to the screening system**. Criterion measures that come from the same “family” or suite of tools are not external to the system. The TRC encourages vendors to select criterion measures and recommends choosing other, similar measures that are on the Tools Chart. An internal measure is considered only if it is paired with an external measure; the vendor must describe provisions that address limitations, such as possible method variance or overlap of item samples.

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

The TRC encourages vendors include data disaggregated by demographic subgroups. Any submission that includes disaggregated data will have a superscript “d” notation on the Tools Chart, and users can access the detailed information by clicking on the cell. Disaggregated data will not be rated; rather, they will be made available to users. A forthcoming advanced search function for the chart also will enable users to quickly locate tools with data disaggregated for the subgroups they are interested in.

6. 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 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 to demonstrate 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 et al., 1991). Measurement models of latent traits (e.g., IRT, confirmatory factor analysis, structural equation models for categorical data) are better suited to provide rigorous examinations of item versus person properties. Speeded tests present additional complications, but they do not remove the need to understand the 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 (Jöreskog, 1979; Vandenberg & Lance, 2000). Most analyses of group differences are simplifications or restrictions on this general model. The TRC will consider any of the following methods as acceptable evidence for bias analysis:

- **Multiple-Group Confirmatory Factor Models for Categorical Item Response** (Meredith & Teresi, 2006): Categorical confirmatory factor analysis 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:** These models include multiple-indicators, multiple-causes (MIMIC; Muthén, 1988; Woods, 2009), or explanatory IRT with group predictors (De Boeck & Wilson, 2004; Van den Noortgate et al., 2003).
  - MIMIC models attempt to test the equivalence of item parameters, conditional on background characteristics or group membership (analogous to an analysis of covariance but for a factor model). Most forms of a MIMIC model represent a restriction of a multiple group confirmatory factor analysis.
  - 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 nonsignificance) of item or person difference parameters.
- **Differential Item Functioning (DIF) From 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 also might consider referencing Meade's taxonomy of standardized effect sizes for DIF that allow for the interpretation of the practical impact of DIF (Meade, 2010).

7. 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 concept is particularly important within problem-solving frameworks when progress monitoring data inform 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 features 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) of the amount of change that occurred, and (c) of the reliability of the change using either statistical or visual methods (Chafouleas et al., 2012; Maggin & Bruhn, 2017). The TRC considers sensitivity to change a unique concept from response to intervention, although the TRC acknowledges that the burgeoning nature of the construct and related methods require some flexibility for documentation. As such, the TRC will currently accept evidence of sensitivity to change based on individual responses to intervention if computation of the metrics is from idiographically collected data. Under the current guidelines, researchers have several methods at their disposal to document sensitivity to change: (a) single measure methods or (b) comparative methods. Descriptions of these broad categories, as well as the specific methods that fall within each, are as follows:

- **Single measure methods** document a particular measure's sensitivity to change. Each of the following 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; it is based on individual responding.
    - **Change metrics.** Change metrics express change in a variable across time (Gresham, 2005; Olive & Smith, 2005). Gresham et al., (2010) recommended several metrics to document change sensitivity in progress monitoring instruments, including absolute change, percentage of nonoverlapping data, percentage change, computation of effect size measures, and the reliability change index. Other statistics may be in this group as well, including alternative nonoverlap statistics (Parker et al., 2011) and regression-based techniques, among others (e.g., Valentine et al., 2016). Computation of these metrics is collected idiographically and compares a student's response to different conditions. Typically, these conditions include a baseline and intervention phase, though conditions also might refer to natural modifications to the environment if there is careful documentation. Sensitivity to change is demonstrated if the metrics document observable change in students 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 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 et al., 2016). Underused 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). Dynamic modeling for evaluating an instrument's sensitivity requires the collection of many data points for individual participants, so many vendors might not be able to use this approach. However, it is an option given its appropriateness for the task. Several dynamic modeling approaches are available, including the traditional *p*-technique, 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 give a rationale for its use.

- **Comparative methods** 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 a more stringent form of sensitivity to change evidence.
  - **Visual analysis.** Miller et al. (2017) presented 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. Miller et al., 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 provided evidence of incremental variability across sessions and allowed visual analysts to determine if the level, trend, and variability across sessions was consistent between the instruments. Sensitivity to change is supported when the instruments represent similar patterns in the data.
  - **Correlational analysis.** Combined with the change metric approach, correlational analyses can evaluate the extent to which change documented through one measure correlates with change documented through another measure. Chafouleas et al., (2012) provided an example of such an approach. Within this study, the researchers calculated two absolute change scores for all 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 the DBR-SIS scores, whereas the second corresponded to change in the SDO scores. Spearman’s rho coefficients were then calculated to examine the extent to which these two sets of absolute change scores correlated with each other.
  - Although less commonly used, **multilevel modeling** also affords a method to compare multiple methods in terms of documented change. Specifically, multivariate growth models can 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 across time.

The TRC acknowledges that there is no accepted framework for documenting sensitivity to change, and the selection of methods will require vendors to consider issues related to the instrument’s construction, scoring rubric, and purpose. Vendors have leeway to select the methods most appropriate for their instrument, although justification for the methods is necessary. TRC members might request additional clarification or metrics if the methods used are inconsistent or unclear.

8. What does the TRC expect vendors to submit for data to support intervention change and 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 the following:

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

9. Can I submit tools that also work 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

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. https://doi.org/10.1177/001440291207800406

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

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

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. https://doi.org/10.1111/bjop.12046

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. 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. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.580.8732&rep=rep1&type=pdf

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*, 364–379. https://doi.org/10.1080/02796015.2010.12087758

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. https://doi.org/10.1207/s15327906mbr4002_3

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. 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). Abt Books.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

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

McDonald, R. P. (2000). A basis for multidimensional IRT. *Applied Psychological Measurement, 24*(2), 99–114. https://doi.org/10.1177/01466210022031552

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. https://doi.org/10.1037/a0018966

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). 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. https://doi.org/10.1037/0003-066X.50.9.741

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. https://doi.org/10.1177/1534508416658007

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). Erlbaum.

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* (pp. 335–344). Sage.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology*, *25*, 313–324. https://doi.org/10.1080/0144341042000301238

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. https://doi.org/10.1177/0145445511399147

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229–244. https://doi.org/10.1177/014662169401800304

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Valentine, J. C., Tanner-Smith, E. E., Pustejovsky, J. E., & Lau, T. S. (2016). Between-case standardized mean difference effect sizes for single-case designs: a primer and tutorial using the scdhlm web application. *Campbell Systematic Reviews*, *12*(1), 1-31. https://doi.org/10.4073/cmdp.2016.1

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. https://doi.org/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. https://doi.org/10.1177/109442810031002

Verhelst, N. D., & Glas, C. A. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, *58*, 395–415. https://doi.org/10.1007/BF02294648

Wang, M., Zhou, L., & Zhang, Z. (2016). Dynamic modeling. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 241–266. https://doi.org/10.1146/annurev-orgpsych-041015-062553

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. https://doi.org/10.1080/00273170802620121

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. https://doi.org/10.1080/15434300701375832

This resource was produced under U.S. Department of Education, Office of Special Education Programs, Award No. H326Q210001. Celia Rosenquist serves as the project officer. The views expressed herein do not necessarily represent the positions or policies of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this document is intended or should be inferred.