Behavior Screening Frequently Asked Questions

1. How does the Technical Review Committee (TRC) consider evidence for screening tools across multiple grade spans and/or have forms for different informants (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. For classification accuracy, the protocol submission form requires that cut points align with students needing behavioral intervention. How does the TRC define students in need of behavioral intervention for this purpose?

Vendors should provide a rationale for how the screener identifies students in need of behavioral intervention. This could include students exhibiting a moderate or high level of risk for the behavior of interest. The TRC uses a consistent definition of students in need of behavioral intervention across all three sets of Tools Charts: screening, progress monitoring, and intervention. For students in need of behavioral intervention, this may include the following: students with an emotional disturbance label; students in an alternative school/classroom; students who demonstrate nonresponse to moderately intensive intervention (e.g., Tier 2); or students who demonstrate severe problem behaviors (e.g., Tier 3), according to an evidence-based tool (e.g., systematic screening tool or direct observation).

3. For classification accuracy, I have data using multiple criterion measures from multiple times of the year. Can I submit all this information?

Yes. The TRC encourages the submission of data using more than one criterion measure and from administrations at different times of the year. Users may be interested in knowing how well a measure predicts risk for more than one outcome, which is why evidence for

more than one criterion measure is useful. Also, in multi-tiered systems of support frameworks, screening involves administration at several time points (i.e., fall, winter, and spring) across the school year, so it is important to understand the degree to which a screener demonstrates classification accuracy at each administration time point. The TRC will rate and report ratings on the Tools Chart for up to six sets of classification accuracy statistics: criterion measure 1 fall administration, criterion measure 1 winter administration, criterion measure 1 spring administration, criterion measure 2 fall administration, criterion measure 2 winter administration, and criterion measure 2 spring administration. The specific criterion measures used will differ for each tool, and the appropriateness of the criterion measure will be factored into the overall classification accuracy rating. Submissions may include data for more than two criterion measures, but they must specify which two measures should be rated. Users will be able to access information on all the criterion measures, as well as the detailed data, by clicking on the appropriate cell in the chart. For time of year, vendors should align the administration time with the closest season (e.g., an October administration would be "fall," a January administration would be "winter"). Regardless of time of year, the TRC requires that at least 3 months pass between the administration of the screening measure and the outcome measure. Vendors do not need to submit classification accuracy data for all six categories; any category for which information is not available will appear on the chart as "—" for "data unavailable."

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

The TRC expects rigorous reliability analyses that are appropriate given the type and purpose of the tool. Regardless of the type of reliability reported, because intended uses for tools can vary, the vendor must provide supporting justification of the choice of emphasis for reliability evidence. Examples of the types of reliability the TRC expects to see submitted include the following:

- **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). Internal consistency is important to report for rating scales that may measure multiple latent constructs.

[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.

- **Test-retest:** Test-retest[2] data should include a justification that explains why the time between test and retest administration is appropriate to the behavior or construct being measured.
- **Interrater:** The TRC requires interrater reliability reports for tests that are subjective and require human judgment (e.g., open-ended questions) versus simple choice selection or computer-recorded responses that would not require 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.
- **Alternate form:** Although not typical for behavior screening, for those tools that have multiple forms (e.g., Form A and Form B), evidence can be provided to 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. (Note: When forms for different raters are available [e.g., teacher, parent], this is not an alternate form because each rater type would be reviewed separately.)

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

[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).

5. What does the TRC expect vendors to submit for validity, 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.

6. For sample representativeness, how are samples classified? What is meant by a cross-validation study?

Sample representativeness refers to the extent to which the samples used to determine the tool's classification accuracy are generalizable to other populations. A tool is more generalizable if studies have been conducted on larger, more representative samples and if cross-validation studies have been conducted.

Samples are classified as either national, regional, or local. A national sample has at least 150 students across at least three of the nine geographical divisions defined by U.S. Census Bureau: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf. A regional sample comes from one or more state samples. A local sample comes from one or more district samples.

Cross-validation is the process of validating the results of one study by performing the same analysis with another sample. In the cross-validation study, cut scores derived from the first study are applied to the administration of the same test and criterion measure with a different sample of students.

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

The TRC encourages vendors to 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 of interest.

8. 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 the 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 do not 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 those complications 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).
- **Differential Test Functioning:** Because classification occurs based on test scores (e.g., fluency, total, IRT based), assessing differential screening at the test level is useful. In examining differential test functioning, vendors might conduct a series of logistic regressions that predict success on an end-of year outcome measure, predicted by risk status as determined by the screening tool, membership in a selected demographic group, and an interaction term between the two variables. Model results that indicate a statistically significant interaction term would suggest differential accuracy in predicting

end-of-year performance for different groups of students based on the risk status determined by the screening assessment (Linn, 1982).

9. Can I submit tools that also work as progress monitoring tools for review by the screening TRC?

Yes, if the tool also can be used for progress monitoring (i.e., the tool can be used for dual purposes). Specifically, the tool must be able to reliably measure change in an overall behavioral domain.

References

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.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360. https://www.jstor.org/stable/1434586

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

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor and W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335–388). Wiley.

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

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.

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

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

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.