Behavior Screening Frequently Asked Questions (FAQ)

1. **How does the TRC consider evidence for screening tools that can be used 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 that is targeted by the screening instrument, in accordance with developer guidelines about target grade spans or ranges (e.g., K-1, K-3). Data must also be reported separately by informant (e.g., teacher, parent, student), if appropriate for the tool. 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. **For classification accuracy, the protocol requires that cut points be aligned with students needing behavioral intervention. How does the TRC define student in needs of behavioral intervention for this purpose?**
   Vendors should provide a rationale for how the screener identifies students in need of behavioral intervention. This could be 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 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. **For classification accuracy, I have data using multiple criterion measures and from multiple times of the year. Can I submit all of this information?**
   Yes. The TRC encourages the submission of data using more than one criterion measure and from administrations at different times of 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 can be useful. Additionally, in MTSS frameworks screening involves administration at several time points (i.e., fall, winter, and spring) across the school year, and it is important to understand the degree to which a screener has demonstrated classification accuracy at each of these administration time points. The TRC will rate and report ratings on the 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 in to the overall classification accuracy rating. Submissions may include data for more than two criterion measures, but must specify which two should be rated. Users will be able to access information on all of the criterion measures, as well as the detailed data, by clicking on the appropriate cell in the chart. For time of year, vendors are asked to align administration time with the closest season (e.g., an October administration would be "fall"

and 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 are not required to submit classification accuracy data for all 6 categories; any category for which information is not available would be noted on the chart as "—" for "data unavailable."

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

The TRC expects reliability analyses that are rigorous and appropriate given the type and purpose of the tool.

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

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

- **Test-retest**: Test-retest* data should be provided with a justification explaining why the time between test and retest administration is appropriate to the behavior or construct being measured.

- **Inter-rater**: The TRC requires 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.

- **Alternate form**: Although not typical for behavior screening, for those tools that do 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 median score of multiple probes) and across time period. Note: When forms for different raters are available (e.g., teacher, parent) this is not considered alternate form as each rater type would be reviewed separately.

Additionally, vendors may submit model-based approaches to reliability. If model-based approaches are used, strong evidence from ***one*** analysis with at least two sources of variance (e.g., time and rater) is acceptable to receive a full bubble. For screening 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 Structural Equation Modeling (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. In evaluating sources of variance, a model-based approach might be founded upon generalizability theory, wherein researchers examine the influence of various screening related facets (e.g., time, rater, screener forms) on the generalizability and dependability of scores.

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.

*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 indicate the assessment's reliability, 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 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 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 requires criterion measures be ***external to the screening system***. Criterion measures that come from the same "family" or suite of tools are not considered to be external to the system.

6. For Sample Representativeness, how are samples classified and 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 considered 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://www.census.gov/geo/reference/gtc/gtc_census_divreg.html*. A regional sample is drawn from one or more state samples. A local sample is drawn 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 that is disaggregated for demographic subgroups (e.g., English learners, students with disabilities, students from 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. A forthcoming advanced search function for the chart will also enable users to quickly locate tools that have data disaggregated for the subgroups they are interested in.

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 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, 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 three 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 Item Response Theory (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 analysis of covariance [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 for interpretation of the practical impact of DIF.

- *Differential Test Functioning*. Given that classification occurs on the basis of test scores (e.g., fluency, total, IRT based), assessing differential screening at the test level can be useful. In examining differential test functioning, vendors might conduct a series of logistic regressions predicting 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 existed for different groups of students based on the risk status determined by the screening assessment (Linn, 1982).

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

Yes, if the tool can also be used for progress monitoring purposes (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.

Appendix A. R Code to calculate AUC statistics

The following code provides an example of how to use R to calculate area under the curve (AUC) statistics. Note that for this example, the *Social, Academic, and Emotional Behavior Risk Screener* (SAEBRS) is being compared to the *Behavioral and Emotional Screening System* (BESS), where the former is the predictor and the latter is the criterion. All analyses are conducted using the pROC package (Robin et al., 2011), which can be used to conduct receiver operating characteristic (ROC) curves.

#Activate the 'pROC' package

```
library(pROC)
```

#Load the dataset. Here, we are choosing to call the data "dat1." We are also telling R where the data file (i.e., "SAEBRS_data.csv") is in the Research folder within the Admin user account.

```
dat1 <- read.csv('/Users/Admin/Research/SAEBRS_data.csv')
```

#Next, we tell R to create an object, which represents the pairing of the predictor and outcome variable of interest. In the example below, we are telling R that we want to run a ROC curve analysis while considering the SAEBRS Total Behavior scale ("SAEBRS_TB") the predictor and the BESS ("BESS_Risk") the criterion. Note that here, the SAEBRS variable is continuous (i.e., summed values ranging from 0-57), whereas the BESS variable is dichotomous (i.e., 0 = No Risk and 1 = Risk).

```
TB<-roc(BESS_Risk~SAEBRS_TB, dat1)
```

#The command below can be used to compute the area under the curve (AUC) for the previously described ROC curve analysis.

```
auc(TB)
```

References

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

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

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*, 837–845.

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.

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

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.

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*, 335-388.

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.

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.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 1-8.

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.