Academic Screening Tool Chart Review: Frequently Asked Questions (FAQ)

1. How does the TRC consider evidence for tools that can be used at multiple grade levels?

Submissions must report data separately for each grade level that is targeted by the screening instrument (in accordance with developer guidelines about target grades). Evidence will be rated and reported on the chart separately for each grade. In cases where data are not available for one or more grade(s) that fall within the grade span targeted by the tool, the TRC will indicate data were not available by including a "—" symbol on the Tools Charts.

2. For classification accuracy, the protocol requires that cut points be aligned with students needing intensive intervention. How does the TRC define a student in need of intensive intervention for this purpose?

For the screening tools chart, the TRC's goal is to review tools that have evidence of the ability to identify students in need of intensive intervention. Therefore, the TRC requires a cut point between the 10th and 20th percentile on the outcome measure.

3. For classification accuracy, I have data for cut points aligned with multiple risk levels (not just intensive intervention). Can I submit this information?

The ratings displayed on the tools chart refer to data drawn from analyses using cut-points aligned with students needing intensive intervention. However, on the third tab of the tools chart (called "Usability Features"), a column is available to indicate the full range of decision rules that the tool covers (e.g., moderate or intensive level of risk), as well as a column to indicate whether or not technical data is available for multiple decision rules. Users can click on these cells to find the detailed information about this evidence.

Note: If a state proficiency assessment is used as a criterion measure, the vendor should not use proficiency levels as the cut point, but rather identify a different cut point aligned with intensive intervention.

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

You may submit information on only one criterion measure for each grade and time of year combination (e.g., Grade 1 Fall). However, you may use different criterion measures across grade and time of year combinations. The TRC will rate and report ratings on the chart for up to three sets of classification accuracy statistics for each grade level: fall, winter, and spring administration of the criterion measure. 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. 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"). Vendors are not required to submit classification accuracy data for all 3 times of year; any time of year for which information is not available would be noted on the chart as "N/A" for "not applicable."

5. 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. If model-based evidence is being submitted for any reliability or validity standard, note that providing Test Information Function (TIF) / Standard Error (SE) plots to judge the relative precision of the model-based estimate(s) is acceptable in place of providing confidence intervals.

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

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

should consider reporting marginal reliability as well as an ability-conditional estimate (e.g., report reliability estimates for students with differing levels of ability to demonstrate evidence that the tool has a sufficient number of items to reliably assess students at various levels of ability; 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 Acceptable Forms of Reliability:

- **Alternate form**: For tools that multiple forms (e.g., fall/winter/spring benchmark materials), 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 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).

- **Inter-rater**: The TRC strongly recommends that inter-rater reliability be reported for tests that are subjective and require human judgment (e.g., open-ended questions, narrative retell) as opposed to simple choice selection or computer recorded responses that would not require inter-rater reliability. The analyses should acknowledge that raters can differ not only in consistency, but also in level. Possible analyses include multilevel models of ratings within judges and students, generalizability theory, and invariance testing in SEM.

*Note that the TRC does not 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).

7. 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 submit a set of validity analyses that offer theoretical and empirical justification for the relationship between the screener 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 that are ***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. The TRC encourages vendors to select criterion measures, and recommends choosing other, similar measures that are on the tools chart. An internal measure is only considered if paired with an external measure and the vendor must describe provisions that have been taken to address limitations such as possible method variance or overlap of item samples.

8. For Sample Representativeness, how are samples classified and what is meant by a cross-validation study and why is this important?

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 (see Appendix B for states by division). A regional sample is drawn from one or more state samples. A local sample is drawn from one or more district samples.

Sample characteristics must include size, date of collection, and location. The TRC encourages vendors to provide demographic characteristics based on the sample but when that is not possible, the TRC will accept demographic characteristics of the school district from which the sample was drawn.

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. Cross-validation is important for understand the degree to which a test can be generalizable to a larger population.

9. How does the TRC consider evidence that is disaggregated for demographic subgroups (e.g., English learners, students with disabilities, 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 they are interested in.

10. 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).
- 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).

Appendix A. U.S. Census Bureau Divisions

Division 1: New England

- Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont

Division 2: Middle Atlantic

- New Jersey, New York, Pennsylvania

Division 3: East North Central

- Illinois, Indiana, Michigan, Ohio, Wisconsin

Division 4: West North Central

- Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota

Division 5: South Atlantic

- Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Caroline, Virginia, West Virginia

Division 6: East South Central

- Alabama, Kentucky, Mississippi, Tennessee

Division 7: West South Central

- Arkansas, Louisiana, Oklahoma, Texas

Division 8: Mountain

- Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming

Division 9: Pacific

- Alaska, California, Hawaii, Oregon, Washington

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. https://doi.org/10.2307/2531595

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.

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

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

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. 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*, 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. 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). 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. https://doi.org/10.1037/0003-066X.50.9.741l

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.

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