Academic Screening Frequently Asked Questions (FAQ)

1. How does the TRC consider evidence for tools that can be used at multiple grade levels? 2
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 needs of intensive intervention for this purpose? 2
3. For classification accuracy, I have data for cut points aligned with multiple risk levels (not just intensive intervention). Can I submit this information? 2
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? 2
5. What does the TRC consider sufficient with respect to sample size? 3
6. What does TRC expect vendors to submit for reliability, and what factors are considered when rating the quality of this evidence? 3
7. What does TRC expect vendors to submit for validity, and what factors are considered when rating the quality of this evidence? 4
8. For Sample Representativeness, what is meant by a cross-validation study and why is this important? 5
9. For Sample Representativeness, what does the TRC mean by "region"? 5
10. How does the TRC consider evidence that is disaggregated for demographic subgroups (e.g., English learners, students with disabilities, students from different racial/ethnic groups)? 5
11. What kind of evidence does the TRC expect to see for Bias Analysis? 5
12. Can I submit tools that can be used as progress monitoring tools for review by the screening TRC? 7

Appendix A. Guidance on preparing classification accuracy evidence 8

Appendix B. R Code to calculate AUC statistics and associated 95% confidence intervals 9

References 10

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 are 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 grades that fall within the grade span targeted by the tool, the TRC will give a rating of "—" to indicate "data were not available."

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 needs 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 TCR expects to see a low cut point and has identified the 20th percentile as a maximum.

   For more information about how to prepare and present classification accuracy data in your submission, see Appendix A.

3. **For classification accuracy, I have data for cut points aligned with multiple risk levels (not just intensive intervention). Can I submit this information?**
   Yes. Although the ratings displayed on the tools chart refer to data drawn from analyses using cut-points aligned with students needing intensive intervention, submissions may include data drawn from analyses using other cut points, representing lower risk levels. These data will not be rated but will be displayed in the supporting detail section of the chart for users to view. 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 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?**
   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 into 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 "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.

See Appendix B for an example of how to use R to calculate area under the curve (AUC) statistics and associated 95% 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) so that the strength of IRT reporting can be fully leveraged in reporting (Green, Bock, Humphreys, Linn, & Reckase, 1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360. Note that for marginal reliabilities, coefficients may not differ much from Cronbach's alpha and can therefore be interpreted using the same guidelines.

If model-based approaches are not used, it is expected that strong evidence for at least two other forms (see list of examples below) of appropriately justified reliability are provided to receive a full bubble. Regardless of the type of reliability reported, given that intended uses

for tools can vary, it is incumbent on the vendor to provide supporting justification of choice of emphasis for reliability evidence.

Examples of Forms of Reliability:

- Alternate form: For 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 across time period.

- Internal consistency (alpha, split-half): Ad hoc methods for item-based measures include internal consistency methods such as alpha and split half. Split half methods are arbitrary and potentially artefactual. Alpha is the mean of all possible split halves (Cronbach, 1951). However, alpha is not an index of test homogeneity or quality per se (Schmitt, 1996; Sijtsma, 2009).

- Test-retest: Test-retest data should be provided with a minimal time period of 1 week (no more than two).

- Inter-rater: 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 offer a set of validity analyses that offer theoretical and empirical justification for the relationship between its tool and a related criterion measure. In other words, the vendor needs to specify the expected relationship between the tool and a criterion, and then use an appropriate empirical analysis to test this relationship. The TRC discourages vendors from providing a large list of validity coefficients correlating with multiple criterion measures, and instead recommends a few analyses that have a theoretical basis about a relationship between the tool and a small set of appropriate criterion measures.

Types of validity may include: evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and/or evidence based on consequences of testing. The vendor may include evidence of convergent and discriminant validity. However, regardless of the type of validity reported, the vendor must include a justification demonstrating how these data taken together demonstrate expected relationships between the measure and relevant external criterion variables. If appropriate, the vendor should take into account the fact that analyses against more proximal outcomes might be expected to show higher correlations than analyses against distal measures, and offer explanations of why this is the case.

It is important to note that to support validity, the TRC prefers and strongly encourages criterion measures that are ***external to the 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. If it is necessary to use internal measures, the vendor must describe provisions that have been taken to address limitations such as possible method variance or overlap of item samples.

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

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. For Sample Representativeness, what does the TRC mean by "division"?

The TRC defines divisions in accordance with the U.S. Census Bureau geographical divisions (see https://www.census.gov/geo/reference/gtc/gtc_census_divreg.html).

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

The TRC encourages vendors to include data disaggregated for demographic subgroups. Any submission that includes disaggregated data will be noted on the chart with a "d" superscript, and users can access the detailed information by clicking on the cell. Note that disaggregated data will not be rated, but instead just made available to users. An advanced search function for the chart will also enable users to quickly locate tools that have data disaggregated for the subgroups they are interested in.

11. 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).

12. 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. Specifically, the tool must be able to reliably measure growth in an overall academic domain over a period of 20 weeks or more. For tools which measure narrower skills that can be mastered in shorter periods of time and that predict competency in broader academic domains (e.g., letter naming fluency, oral counting), the TRC recommends submitting for review by the screening TRC, to be consistent with the way in which users are most likely to use the tool.

Appendix A. Guidance on preparing classification accuracy evidence

Classification Accuracy will be rated separately for each criterion measure and time of year for the administration (e.g., Fall, Winter, Spring). Ratings will be provided for up to two different criterion measures and up to three different time points. Data for additional criterion measures or administration times may be reported but will not be rated.

Vendors will need some specific data points to calculate the indices of classification accuracy that must be reported in the protocol. The matrix below shows these data points and their relationship to each another.

	Students Actually "At-Risk"	Students Actually "Not At-Risk"	Total
Students Classified as "At-Risk"	True Positive a	False Positive b	a + b
Students Classified as "Not At-Risk"	False Negative c	True Negative d	c + d
			N = a+b+c+d

Filling out the Classification Accuracy section of the screening protocol will require you to use the following formulas:

False Positive Rate = b/(b+d)

False Negative Rate = c/(a+c)

Sensitivity = a/(a+c)

Specificity = d/(b+d)

Positive Predictive Power = a/(a+b)

Negative Predictive Power = d/(c+d)

Overall Classification Accuracy Rate = (a+d)/(a+b+c+d)

For assistance with calculating area under the curve statistics, see Appendix B.

Appendix B. R Code to calculate AUC statistics and associated 95% confidence intervals

The following code provides an example of how to use R to calculate area under the curve (AUC) statistics and associated 95% confidence intervals. Note that for this example, the *AIMSweb* is being compared to the *Pennsylvania State Assessment*, 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., "AIMSweb_data.csv") is in the Research folder within the Admin user account.

```
dat1 <- read.csv('/Users/Admin/Research/AIMSweb_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 AIMSweb R-CBM scale ("AIMSweb_RCBM") the predictor and the Pennsylvania State Assessment ("Penn_Risk") the criterion. Note that here, the AIMSweb variable is continuous (i.e., summed values ranging from 0-57), whereas the Pennsylvania State Assessment variable is dichotomous (i.e., 0 = No Risk and 1 = Risk).

```
TB<-roc(Penn_Risk~AIMSweb_RCBM, dat1)
```

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

```
auc(TB)
```

#This final command can be used to compute the 95% confidence interval around the AUC using the DeLong, DeLong, and Clarke-Pearson (1988) asymptotic exact method.

```
ci.auc(TB, conf.level=0.95)
```

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.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

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.