**Issues related to comparing new and old assessments and comparing trends (linking & establishing trends)**
**(Created by MARC team, November 20, 2013)**

| **Issues** | **Considerations** |
|---|---|
| **Linking tests for proper comparisons**<br><br>• Beginning in the fall of 2014, according to new standards set forth by the Common Core State Standards initiative (CCSS), the state of Maryland, under PARCC, will replace MSA with PARCC assessment, which will differ in:<br>(1) Content coverage<br>(2) Scope and sequence<br>(3) Standards<br>(4) Psychometric properties<br>(5) Score distributions<br>(6) Difficulty level<br>(7) Levels of competency<br>• MSA has 3 levels (Basic, Proficient, Advanced)<br>• PARCC assessment reports 5 levels<br>(8) Cut scores for each level<br><br>• Such discrepancies between new and old assessments require a proper linking study in order to capture student progress.<br><br>**Lack of linking causes misinterpretations**<br>Linking state performance standards to NAEP scale (USDE 2011):<br><br>• Studies show that many states are setting the bar significantly lower in elementary school than in middle school, giving parents, educators, and the public the false impression that younger students are on track for future success.<br><br>• Improvements in passing rates on state tests after NCLB’s enactment can largely be explained by declines in the difficulty of those tests.<br><br>• Diversity of test difficulty and performance level descriptions can make comparisons difficult. | **Linking designs**<br><br>**1. Common-item linking method**<br>• *Description*: Anchor items are embedded into the new test form. Test forms contain non-anchor items as well as anchor items. Anchor items could be internal and external. The internal anchor items will be used in scoring students while the external anchor items will not be used for scoring students, rather just for the purpose of linking the two tests.<br>• Administer a subset of MSA questions in addition to new PARCC tests<br>• *Advantages:*<br>1. It is the least resource-intensive since it requires only one test administration.<br>• *Disadvantages*<br>1. When curriculum and instruction have fully migrated to match the new standards, student performance might not be the same (Kirkpatrick, Turhan & Lin, 2012).<br>2. Building anchor sets might be challenging because new and old administrations might have different test specifications.<br>3. Anchor items must maintain stability across tests which were constructed according to different test blueprints.<br><br>**2. Common person or single group linking method**<br>• *Description:* Same subjects take two tests on different occasions. Tests are designed to the same specification, but share no common items. Common persons are used to link the two tests and put them onto the same scale<br><br>• Single group counter-balanced design: Give two forms in different orders to a random sample of the single group to control for fatigue and practice effects. Both subgroups of the single group will take two tests, but in different orders.<br><br>• Administer a drill-and-practice exam comprised of the MSA questions with the same test specifications similar to the PARCC assessment. The single group of students taking part in the linking study will be randomly assigned to each of the two sequences: taking MSA first then PARCC test while the other taking the PARCC first then MSA.  Linking between MSA and |

| | |
|---|---|
| • Simply looking at proportions of students classified as ‘proficient’ is insufficient to judge the effectiveness of an educational system because proficiency is defined differently by different tests at different grades and for different subjects. | PARCC tests will be set up based on the common persons.<br>• *Advantages:*<br>1. No extra efforts are needed to construct an anchor form of a MSA test.<br>2. Helpful in providing evidence of construct validity which is not provided by common-item approach<br>• *Disadvantages:*<br>1. Requires two administrations<br>2. Hard to control for practice and fatigue<br>3. Testing time is doubled<br>**3. External test linking method**<br>• *Description:* Use a separate, intermediary test to link two tests that have different scales. For example, students who take MSA, PARCC, and the Measures of Academic Progress (MAP) by NWEA could be used to set up the chained linkage of MSA and PARCC on the same scale via MAP.<br>• *Option 1:* Equipercentile linking method: Use equipercentile linking method to link MSA to MAP, then use equipercentile linking method again to link MAP to PARCC. Ultimately, using double linkage, MSA and PARCC tests are linked.<br>• *Advantages:*<br>1. Methodologically defensible.<br>2. No requirement of extra testing time.<br>3. Produces cut score estimates and pass/fail predictions that are essentially equivalent to more complex methods (Cronin, et al., 2007).<br>• *Disadvantages:*<br>1. Assume representative samples used in equipercentile equating<br>2. Requires three test administration test data are available.<br>3. Larger equating errors are likely<br>• *Option 2.* Fixed item parameter method: given one test is used to construct a base scale; the item parameters estimated from the base scale are fixed to put the external test onto the same base scale. Then the item parameters for the external test are fixed in a concurrent calibration to put a third test onto the base scale.<br>• Use the PARCC test to construct the base scale. Do concurrent calibration of PARCC and MAP tests by fixing the item parameters of PARCC test. |

| | |
|---|---|
| | Then, fix recalibrated item parameters of the MAP test which are on the scale of PARCC test and concurrently calibrate MAP and MSA. MSA tests are now on the PARCC scale. If MSA scale is used as the base scale, the process is reversed to put the PARCC test onto the MSA scale.<br>• *Advantages*:<br>1. Methodologically easy to implement as long as the item response data and item parameter estimates from the base scale are available for students taking the three tests at different time points.<br>2. More flexible approach in comparing two tests once the two tests are put on the same scale.<br>• *Disadvantages*:<br>1. Efforts need to be made to locate item responses data from three tests: MSA, MAP, and PARCC and have the item parameters available for the base scale test.<br>2. Small sample size would be an issue<br><br>**Key points from linking study in recent NAEP and TIMSS comparison study (NAEP, TIMSS 2013). (**Braided booklets containing NAEP and TIMSS items were administered as the common items)<br><br>**1. Statistical moderation**: Aligns score distributions from one test with the other so that mean and variance are matched and comparisons can be made. Classical linear equating method could be used to achieve comparable score distributions across tests.<br>**2. Statistical projection**: Developing a function used to predict a score on a separate test from the observed test. This is essentially a regression prediction, not really making scores from two tests comparable.<br>**3. Calibration:** Uses item response theory to calibrate items directly onto a different test's scale by using a braided booklet with items from both tests.<br>*(Note: In this study, statistical moderation was favored since it required the least amount of computation)* |
| **Comparing trends across different tests**<br>The new PARCC assessments will inevitably disrupt trends. In response, trends arising from the new assessments will need to be compared to different trends that use different tests but have similar content. For example, National Assessment of Education Progress (NAEP) research studies show that most states' proficiency standards are at or below NAEP's definition of *Basic* performance. (example: In grade 4 reading, 35 of the 50 states set standards for proficiency (as measured on the | **Some popular methods for comparing trends across tests**<br>**(1) Change in Proportion Above Cut-score (PAC)**<br>• Looks at the change in percentage of proficient students above a certain cut-score, after a proper linking study.<br>• Follows logic that changes should be the same across test measuring the same achievement for the same students<br>• PAC statistics will be variant over different cut-scores |

| | |
|---|---|
| NAEP scale) that were lower than the scale score for *Basic* performance on NAEP and another 15 were in the NAEP *Basic* range). Such discrepancies call for a proper trend comparison study across different test. | **(2) Effect sizes**<br>• Expresses trend differences in averages in terms of standard deviation units<br>• These standard deviations are obtained from a pool of standard deviations of the distributions of trends to be compared (Hedges & Olkin, 1985)<br>**(3) Transition models**<br>• Students are classified into categories based on statewide definitions of basic, proficient and advanced in each grade<br>• Each category is divided again into more subcategories.<br>• Students are followed into the next grade level and their performance is observed.<br>• Change scores are calculated for each student to see trends.<br>**(4) Other methods**<br>• Ho (2009) suggests non-parametric graphs and statistics to compare trends and gap-trends across different tests, because they are more stable under different scales.<br>• Probability-Probability plots (Wilk & Gnanadesikan, 1968) have been proposed for the use in the context of gaps on educational tests by Haertel, Thrash, & Wiley (1978), Spencer (1983b), and Livingston (2006)<br>*(Note: These methods were tested using NAEP trend data and state tests trend data)* |

**References**

Bandeira de Mello, V., Blankenship, C., and McLaughlin, D.H. (2009). Mapping State Proficiency Standards Onto NAEP Scales: 2005–2007 (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. G. (2007). *The proficiency illusion*: Thomas R. Fordham Institute and Northwest Evaluation Association.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. Journal of Educational and Behavioral Statistics, 34, 201-228.

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test score distributions. Journal of Educational and Behavioral Statistics, 27, 3-17.

Livingston, S. A. (2006). Double P-P plots for comparing differences between two groups. Journal of Educational and Behavioral Statistics, 31, 431-435.

NWEA (2010). California linking study: A study of the alignment of the NWEA RIT scale with the California Standardized Testing and Reporting (STAR) program.

Schafer, W. D., Lissittz, R. W., Zhu, X., Zhang, Y. (2012). Evaluating teachers and schools using growth models. *Practical Assessment research and evaluation, 17(17),* retrieved from http://pareonline.net/pdf/v17n17.pdf

Spencer, B. D. (1983a). Test scores as social statistics: Comparing distributions. *Journal of Educational Statistics*, *8*, 249-269.

Spencer, B. D. (1983b). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement*, *20*, 317-333.

USDE (2011). Mapping state proficiency standards onto the NAEP scales: Variation and change in State Standards for reading and mathematics, 2005-2009.

Renaissance Learning (2013). Pathway to proficiency: Linking the STAR reading and STAR math scales with performance levels on the New York State Assessment Program (NYSTP) for English language arts and mathematics.

Kirkpatrick, R., Turhan, A., & Lin, J. (2012). Linking two assessment systems using common-item IRT method and equipercentile linking method. Paper presented at the Annual meeting of the National Council of Measurement in Education, Vancouver, Canada.

Yu, C.H., & Popp, S.E. (2005). Test equating by common items and common subjects: concepts and applications. *Practical Assessment, Research & Evaluation*, *10(4)*. Retrieved from http://pareonline.net/pdf/v10n4.pdf