TEST ASSESSING SECONDARY COMPLETION™

The Comparability of Computer and Paper-Based Forms in Light of Technology Enhanced Items

The use of technology enhanced items for state accountability testing is increasing and expected to be the norm by 2015. Thus, the TASC Test Assessing Secondary Completion™ is adopting the use of technology enhanced items to maintain comparability to the next generation of high school assessments. The use of technology enhanced items has been driven based on the belief that they can support "better" measurement of certain standards. We note that this is analogous to the use of constructed-response items to measure certain standards "better" than selected-response items—a transition that occurred in the 1990's.

Background

Data Recognition Corporation has decades of experience developing comparable test forms containing different item types that are administered interchangeably and validly report student scores on a common scale. DRC developed a national standardized assessment system, *TerraNova*®, in 1996 which consisted of multiple editions. The Survey and Complete Battery editions consisted of selected-response items only. The Multiple Assessments edition consisted of a mix of selected-response (items which were also on the survey edition) and constructed-response items. DRC developed a single scale for the multiple editions by conducting a standardization study in which a large sample of nationally representative students (the norm group) took various standardization forms. Data from the standardization study were used to place all items on the same scale.

Item quality and the validity of scaling multiple item types together were assessed through classical and item response theory (IRT) analyses. In particular, item quality was assessed through the examination of p-values and distractor analyses. The validity of scaling the multiple item types together was assessed by examining (a) appropriate item-to-test correlations (a broad measure of how the items measure the same construct) and (b) item response theory item fit statistics. The elimination of items with poor classical and IRT model fit meant that the items remaining in the pool measured the common construct appropriately. The use of items with good model fit meant that forms developed based the construct structure used for the calibration would provide comparable scores—students would be expected to obtain the same scale score regardless of the form administered.

To support the selection of final forms that could be used interchangeably, forms for each edition were assembled that adhered to a common test blueprint in terms of content structure, such that the resulting test characteristic curves appropriately overlapped the same target curve. It should be noted that different forms of the same edition, say forms A and B Survey or Forms A and B Multiple Assessments, measure the same standards with different sets of items, but adherence to the same standards structure and matching test characteristic curves supports the comparability of scores reported from the different forms. The resulting forms of *TerraNova* for the multiple editions have been used for over a decade to support the reporting of valid student norm and criterion referenced scores for moderate and high stakes purposes.

Creating Comparable TASC Test Forms

The same techniques described above are applied to the development of the TASC test computer and paper-based forms. We develop a common test blueprint in terms of content structure and create

sufficient pools of items, while concurrently scaling them. Then we eliminate items that do not perform well based on the criteria described above, and develop forms for both testing modes that match the designated content structure and test characteristic curve targets. This approach supports the reporting of valid student scores to support the intended uses of the TASC test.

Content Approaches for Developing Computer and Paper-Based Items Measuring the Same Standards
Technology-enhanced (TE) items administered via computer and paper-based items that are comparable in terms of standards alignment and measure the same construct should provide examinees an opportunity to demonstrate evidence of the same content knowledge and skill(s). To support such comparability DRC will use multiple approaches. First, when appropriate, DRC develops pairs of items that use the same stimuli and are alike or similar in length, reading complexity, and cognitive complexity. However, in some cases the nature of the TE item does not allow the close adherence of these facets of the item in the development of items that measure the same standards. In these cases, new paper-based items are developed that measure the same standards as TE items. These items are developed to support the assembly of comparable forms of the TASC test that maintain the same content structure and which undergo the same psychometric methods to support the comparability of the different forms.

The following examples provide a demonstration of the methods used to support comparability across item types:

Math
Consider the following standard: Add and subtract 3 digit whole numbers with and without regrouping. In this case the two comparable TE and paper-based items must match the target operations (addition or subtraction) and the regrouping requirements (with and without regrouping). If possible, both TE and paper-based items are developed to match the number of times regrouping occurs. If the TE item requires a graphic to solve the item, then the paper-based comparable item is developed with an appropriate graphic.

Reading
Consider the following standard: Compare and/or contrast characters in two literary texts. In this case the two comparable TE and paper-based items must match in task (compare and/or contrast) and the number of texts associated with the items (2 or 3 literary texts). Because this standard indicates a text dependency, the best solution to obtain comparable items is to associate the pair to the same two texts and place the items in proximal locations in their respective test forms. If different texts must be used for the TE and paper-based items then the texts will have similar characteristics including length, reading complexity, genre, and character types so that the TE and paper-based items elicit comparable evidence of learning.

TEST ASSESSING SECONDARY COMPLETION™

Research

Research that will be done on pairings will include:

- Item comparability on each test
- Comparability across forms
- Goal is to make items as similar as possible
- Technology Enhanced Items

TE and TE twins that are placed in the same place are important for:

- Research that will be done on pairings
- Item comparability on each test
- Comparability across forms
- Goal is to make items as similar as possible

How can technology elicit deeper evidence of learning, not just to introduce technology:

- MCR elicits more evidence of learning-will and should elicit more information
- TE items are not all the same but rather comparable

Creating vs. selecting different cognitive demands:

- Overall form results are the same
- Not so concerned that items TE and TE twins are same

For additional information visit TASCTest.com or call DRC at 800.538.9547

DATA RECOGNITION DRC CORPORATION