National Center and State Collaborative

Implications of Post-NCSC Project Scenarios for Future Test Development

Brian Gong

Center for Assessment

Development of this report was supported by a grant from the U.S. Department of Education, Office of Special Education Programs (H373X100002, Project Officer: Susan.Weigert@ed.gov). The contents do not necessarily represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

Executive Summary

NCSC (the National Center and State Collaborative) had as a major deliverable to develop operational-ready assessments in ELA (English language arts) and mathematics, for grades 3-8 and high school for students with severe cognitive disabilities who qualify for an alternate assessment with alternate achievement standards (AA-AAS). In 2015 NCSC produced multiple usable forms for each grade/content area, based on an extensive field-test. It would be possible for a state to use the NCSC 2015 forms for a number of years to support an operational assessment and accountability program. However, other likely future post-NCSC test usage scenarios require varying amounts of new item/test development. This paper examines future test development needs and issues for a number of likely scenarios for users of NCSC assessment materials following the completion of the federal funding of the project.

The scenarios include:

- Using the NCSC assessments without modification
- Using the NCSC assessments with minor modifications
- Developing new forms for ELA and math following the NCSC blueprint
- Changing the blueprint to incorporate Writing prompts and
- Developing a computer-adaptive testing approach without modifications, with new item development, and with Writing prompts

In addition to item/test development, the paper also considers implications for scoring, scaling, standard-setting, and reporting; refinement of item and test specifications; and practical implications for having multiple, flexible user/developer conditions.

Introduction

A major deliverable of the federally funded NCSC project (the National Center and State Collaborative) was to develop operational-ready assessments in ELA (English language arts) and mathematics, for grades 3-8 and high school for students with severe cognitive disabilities who qualify for an alternate assessment with alternate achievement standards (AA-AAS). In 2015 NCSC produced multiple usable forms for each grade/content area, based on an extensive field-test. It would be possible for a state to use the NCSC 2015 forms for a number of years to support an operational assessment and accountability program. However, other likely future post-NCSC test usage scenarios would require varying amounts of item/test development. This paper examines future test development needs and issues for a number of likely scenarios for users of NCSC assessment materials following the completion of the federal funding of the project.

The scenarios include:

- Using the NCSC assessments without modification
- Using the NCSC assessments with minor modifications
- Developing new forms for ELA and math following the NCSC blueprint
- Changing the blueprint to incorporate Writing prompts and
- Developing a computer-adaptive testing approach without modifications, with new item development, and with Writing prompts

In addition to item/test development, the paper also considers

- implications for scoring, scaling, standard-setting, and reporting;
- refinement of item and test specifications; and
- practical implications for having multiple, flexible user/developer conditions.

The paper is organized in the following sections:

A. What has been developed
B. What is needed to be developed for likely future scenarios
C. Discussion of implications of new development

Items and Tests Developed by NCSC

NCSC developed a number of assessment products. These included materials that were planned to be administered in 2015 to produce student scores, items that were intended to be finalized for an item pool ready to be used in future tests but were not used to produce student scores in 2015, and items that were pilot-tested but whose development was not completed in 2015. Each of these is described briefly below. More complete documentation is referenced.

Assessment Materials Used for Student Scores in 2015

Test Forms and Items

NCSC in 2015 developed four test forms at each grade for ELA/Literacy and for mathematics. All four forms within a grade/content area were equated to each other and placed on a common grade-level scale. Each form addressed the NCSC grade/content test blueprint, and consisted of Common, Matrix Core, Matrix field-test items.

The Common items were the same across all the forms, and were intended to be used to produce student scores. In addition, each form included a set of items unique to the form which were also used to produce student scores (Matrix Core items). The "core" items are those that were used to produce student scores for the form, and consisted of the Common and the Matrix Core items. In addition, each form included a set of items unique to the form which were not used to produce student scores (Matrix field-test items). On the basis of the 2015 results, the Matrix field-test items were placed into one of three categories: 1) acceptable for placement on future operational tests, 2) a candidate for field-testing in the future following editing, and 3) dropped from consideration.

The NCSC operational item pool thus consists of Core items and fully field-tested items acceptable for placement on future operational tests. The NCSC developmental item pool consists of items that were field-tested in 2015 and identified as being a candidate for revision and re-field testing. NCSC also has items that were not field-tested in 2015: these include ELA and mathematics items pilot tested prior to 2015 (Pilot Test 1 and Pilot Test 2). In addition, most Writing Prompts should be considered as in a developmental stage.

Other NCSC Materials

NCSC produced comprehensive technical documentation of the assessment materials, describing the history of the project, constructs and claims, test development process and results, psychometric characteristics including validity argument and evidence for the items, forms, scale, achievement levels, and reports. Much of this documentation was the basis for the submission by states of the NCSC assessment to Peer Review in 2016. At the time this paper was written, the results of Peer Review were not available.

In addition, NCSC produced model instructional materials, which are valuable for explicating the constructs and for achieving the project's goals, which were to support improved student achievement.

Implications of Future Use Scenarios for Item/Test Development

Several like scenarios are described below in terms of implications for item/test development:

- Using the NCSC assessments without modification
- Using the NCSC assessments with minor modifications
- Developing new forms for ELA and math following the NCSC blueprint
- Changing the blueprint to incorporate Writing prompts and
- Developing a computer-adaptive testing approach without modifications, with new item development, and with Writing prompts

Using the NCSC Assessments without Modification

The NCSC project in 2015 produced four test forms each for ELA and mathematics in each grade 3-8 and high school. A state could use these forms to administer a viable assessment for a school accountability program for several years. One way to do this would be to administer a different form each year in a non-public order to reduce the possibility of directly teaching to the test. For example, if each form were administered twice, the state would have sufficient forms for eight years, which is longer than most state alternate assessment programs have lasted without major modification.

Using the NCSC Assessments with Minor Modifications

It is likely that some states will want to modify the 2015 forms in minor ways, primarily to strengthen the technical qualities of the assessment forms. For example, the 2015 technical report identifies some items on some forms that made the test less reliable (i.e., had negative point-biserial correlations). To the extent that these weak items could be replaced by existing operational-ready items in the item pool that met the test blueprint, it would be a relatively minor task to develop stronger test forms. In fact, this was done by the MSAA collaborative in 2016, as described in the technical documentation. The MSAA modification exemplified what could be done using only test forms and items that were developed by NCSC in 2015.

Other states may wish to do the same type of modification, since it is relatively simple to do and improves the technical quality of the test forms.

There may be other relatively minor modifications that users may wish to make in the NCSC 2015 materials to make the tests better. Some items may be improved by changing how much scrolling is required, or by modifying the order of the correct answer keys. Some item-level

improvements may be possible with existing items in the operational item pool. However, some item-level improvements may require development of new items.

Developing New ELA and Math Test Forms Following the NCSC Blueprint

Some users may wish to make more substantial changes to the existing test forms. At some point these might be better thought of as new test forms. For example, many state testing programs have in their test blueprints that each annual test will have a certain proportion of new items: the proportion might range from a modest 20% to 100%.

Some reasons a state might wish to have a substantial proportion of new items are for test security, to improve the characteristics of the test, or to produce released items to inform understanding about the test.

Users of the NCSC test materials may not be as concerned with test security as a regular assessment, since the design of the alternate assessment is not highly secure. For example, test administrators often have access to the test for several days prior to administration to allow them to prepare any allowed individual administration supports, and the test administrations conditions may not be highly resistant to compromises by the test administrator unless a proctor is present at all times during administration and recording of the student response. Still, some development of new items may be viewed as prudent. The introduction of new test items consistent with the test blueprint and with the content standards may also help safeguard undesirable narrowing of the curriculum to focus on only the few assessment items on a test form. On the positive side, releasing multiple test items may help educators, parents, and the public understand the test in ways that no verbal descriptions could accomplish.

More extensive improvement of the technical characteristics of the test may be possible by replacing many items. This may be desired, even if the existing test forms are acceptable, to improve the reliability/precision and/or validity of the assessment. For example, some analyses of the 2015 NCSC test forms have pointed out that the items do not have as large a range of difficulty as might be wishes, and some items do not conform to the learning progression theory that was hypothesized for them as "tiers." In such cases, careful item development and test construction may create better tests over time, often without requiring new scaling or standard-setting.

Such substantial modifications to the test blueprint would usually require substantial and on-going item development, test form development, and test equating. Any testing contractor should be able to provide and implement a plan for such activities. However, the development may be limited by the number of students; this is discussed more in the last section.

Changing the Test Blueprint to Incorporate Writing Prompts

The current NCSC test blueprint addresses Writing through a couple of selected-response format items. NCSC designed and developed a number of Writing Prompts which required more extensive and complex student responses. However, because the Writing Prompts were on a different development schedule than the other content areas, a selection of Writing Prompts were field-tested in 2015 but were not used to generate student scores.

If a user wished to incorporate Writing Prompts into an assessment based on the NCSC test materials, the user would need to do several things:

- Decide on the scoring method – the Writing Prompts were developed with a three-trait analytical scoring approach matched to the Writing Prompt item specifications that emphasized structural differences in complexity and support between Writing Prompts of four different "tier" levels. NCSC conducted a special study of the scoring approach for a selection of Writing Prompts in 2015, which built on a previous study that used a different scoring approach. Although the scoring approach appeared viable, as with all analytic trait scoring approaches of writing, the user would need to weigh the advantages and disadvantages between a more complex analytic scoring approach with a holistic scoring approach. One advantage of the analytic approach is that it yields more score points, which helps justify the relatively longer time it takes to administer the Writing Prompt, and it potentially provides a basis for providing more nuanced feedback in reports. However, trait scores often have such limited precision that they do not provide a reliable way to distinguish between most students' performance by trait, nor between different students by trait.
- Decide on the approach to control acceptable variation in difficulty between Writing Prompts. NCSC developed the Writing Prompts with the intent that they would be systematically more difficult by increasing tiers, from tier 1 to tier 4. In addition, the Writing Prompts within each tier should be relatively similar in terms of difficulty. The user will need to decide how to control the relative difficulty of the Writing Prompts, both across tiers and within tiers. This challenge of equating Writing Prompts is common in all assessment programs where the Writing score is based on student responses to one or two prompts, and the prompts differ across years and sometimes within year across forms.
- Decide on how the Writing Prompt will be reported—by creating a separate Writing subscore, or by having only a total ELA score (with perhaps non-scaled information from the Writing Prompt such as raw score or strength/area to improve). If the decision is to create a Writing score, the user will need to decide whether/how to combine the Writing Prompt scores with the scores on the selected response Writing items that were included in 2015 for generating student scores, i.e., the user must decide whether to produce a Writing Prompt score and a Reading/Writing score; a Writing score combining the Writing Prompt and the selected response items; or an overall ELA score without separating Reading/Writing.

- Decide on whether/how to scale ELA. The Writing Prompt performance might be included in an overall ELA scale, or an overall ELA score might be created by a composite of the Reading scale and Writing score.
- Decide on how standard-setting will be conducted. The addition of the Writing Prompt may be viewed as such a change in construct that a new standard-setting in ELA is needed (either total or combined Reading-and-Writing). The decision about the change in construct could be informed by analyses of the relationship between Writing Prompt scores and other components of the 2015 ELA test, as well as by a content analysis of the performance level descriptors (what do the PLDs represent in terms of writing?)..
- Decide on test administration. Where will the Writing Prompt be placed on the test? If the Writing Prompt is incorporated into a fixed form test, then how would it be decided which tier Writing Prompt would be administered to all students? Might some way be devised to adapt the Writing Prompt on some basis other than performance on the state ELA test?

Developing a Computer-Adaptive Testing (CAT) Approach

NCSC was interested in producing a computer-adaptive test (CAT) but due to time and development constraints, was not able to produce an operational version. However, NCSC did sponsor two studies of possible CAT designs. One study was done by CTB earlier in the project prior to large-scale field-test item data being available. A second, extensive study was conducted by Ric Luecht featuring simulations using the NCSC 2015 item data.

The Luecht study examined several possible CAT models, primarily under the condition of using the 2015 NCSC item pool, and then made recommendations regarding future item development.

This section recaps Luecht's findings and recommendations, and then briefly considers the addition of Writing Prompts to a CAT.

CAT Simulation Findings

Luecht found that three CAT models were viable and made substantial improvements in precision over the NCSC 2015 fixed forms. These CAT models could be implemented using the NCSC 2015 item pool of operational items. However, Luecht advocated developing more items in specifically targeted ways that would make the CAT models more precise and useful, especially at the upper and lower ends of the student performance distribution. Developing many items has practical implications, which are discussed in the last section.

Incorporating Writing Prompts into CAT

One advantage of a CAT design is that it would allow Writing Prompts of different intended difficulty to be administered in a targeted way to each individual student based on the student's test performance, unlike on a fixed form test. This would be especially appropriate for the

NCSC Writing Prompts that are designed to be quite different in terms of complexity, difficulty, and construct from tiers 1 through 4.

One way to incorporate a Writing Prompt into a CAT would be to administer the Writing Prompt after information has been gathered on the student's performance to allow some estimate of which level of Writing Prompt difficulty would be more appropriate. If the CAT were a two-stage model, then the Writing Prompt could be administered either following the first stage or the second stage. If the CAT were a three-stage model, then the Writing Prompt should be administered following the second or third stage so that the most information is available to appropriately assign the Writing Prompt. Note that as long as the Writing Prompt response must be human-scored, then the total ELA and any other scores that incorporate the Writing Prompt will not be reported immediately after the conclusion of testing, as would be the case if the test consisted only of machine-scorable items. Of course, the user may wish to wait to issue reports for a variety of reasons (e.g., include state results on the score reports; avoid reactions by educators or others to partial results).

Implications of New Item/Test Development

In this section, several implications of new item/test development are discussed, including:

- Item/Test development materials
- Practical requirements for test development
- Governance and terminology

Item/Test Development Materials

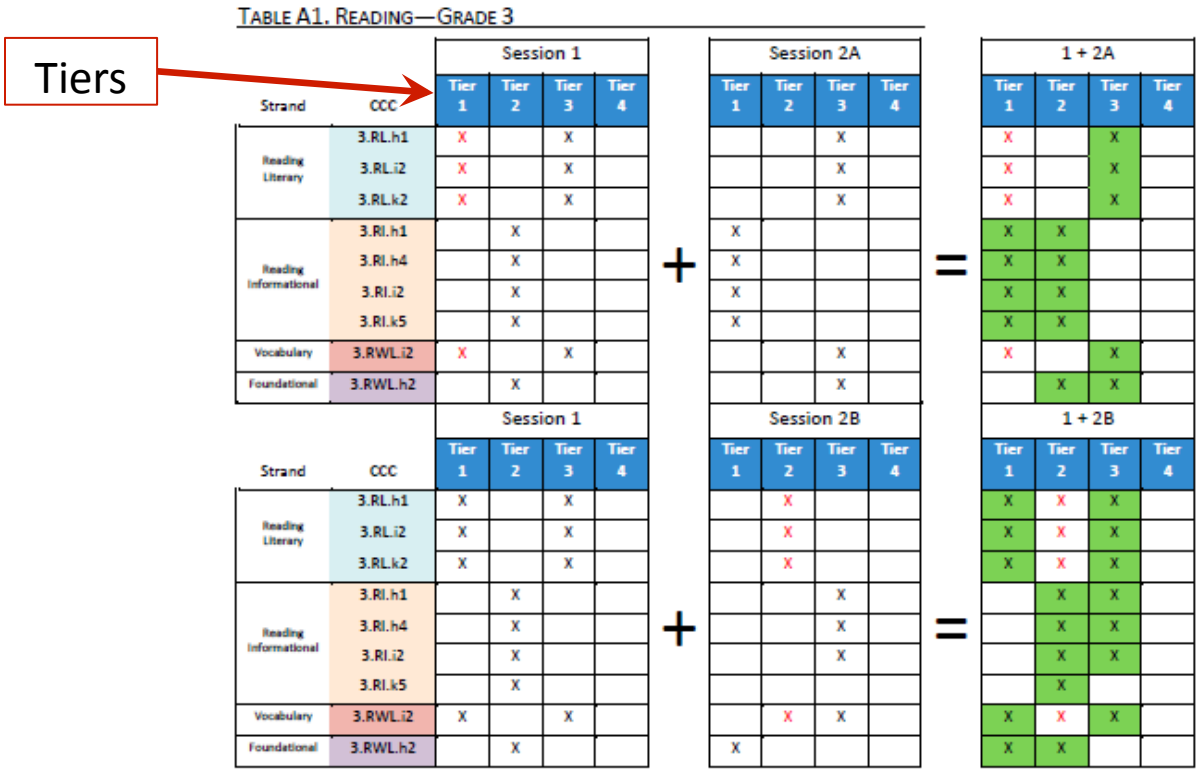
Development of new items and tests should be systematic. NCSC developed tools, including specifications and policies to help it develop items and tests consistently and efficiently. Over the course of the project various modifications were made, informed by each stage. It would be useful for future users to review and make sure the NCSC item and test specifications are internally consistent and reflect the most recent work—or more importantly, reflect the user's intended design.

One particular aspect of NCSC test development that would benefit from being clarified and made consistent at least is the place of "tiers." Tiers were originally posed as distinct specifications. One position was that these were tied to foundational distinctions, such as a learning progression. Another position was that tiers were a useful heuristic for guiding development of items with a range of difficulties, but had no firm theoretical place and should not be learned on too heavily in actual test specifications.

Users should decide how they view tiers, and make the item and test development materials and specifications consistent with that view. For example, Figure 1 shows the test blueprint for

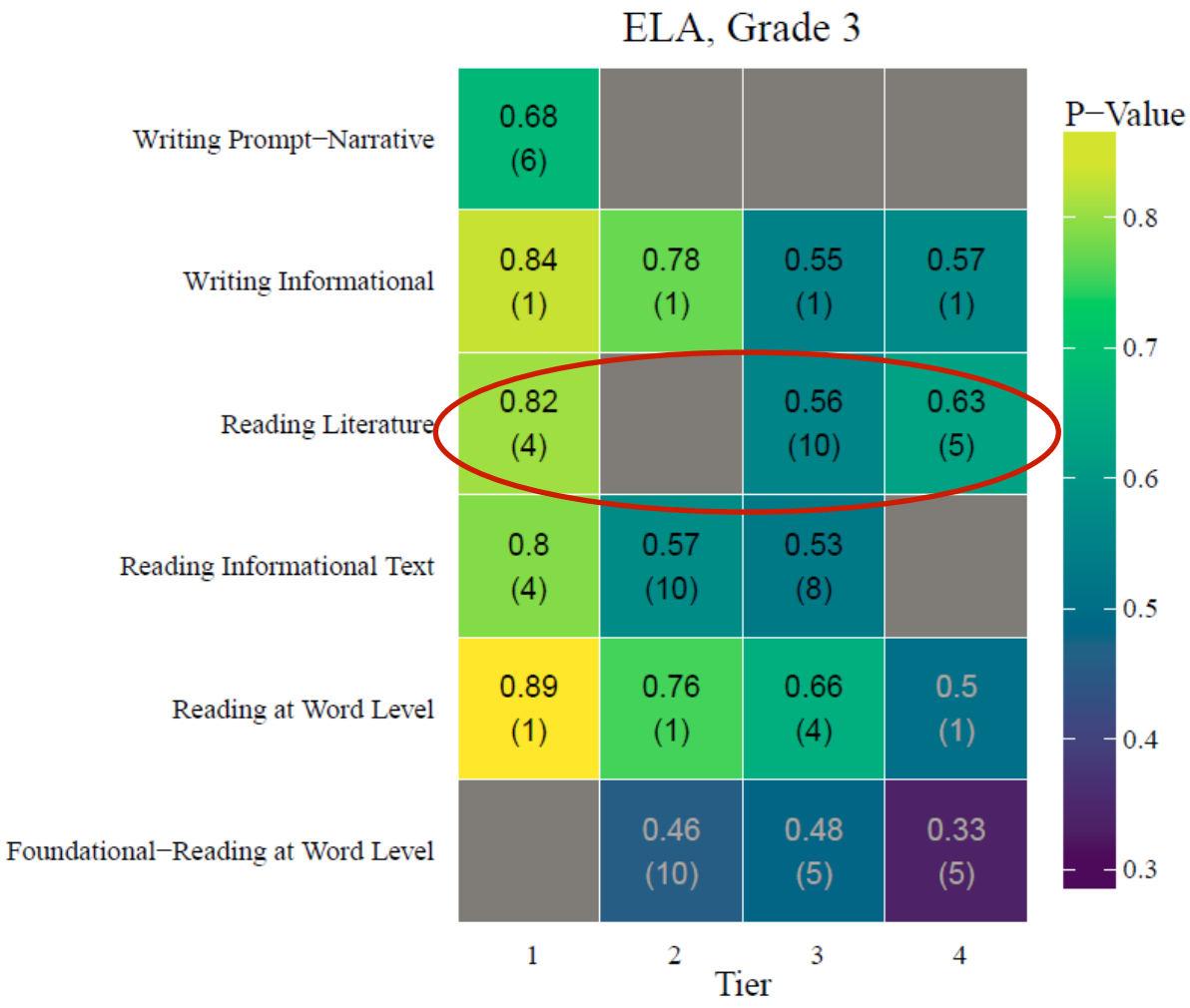
Reading, Grade 3, as drawn from the Technical Manual for the NCSC 2015 tests. Note that the test blueprint is expressed in terms of selecting items by content specifications ("Strand and CCC") and **tiers** (1-4).

Figure 1: Test blueprint, Reading, Grade 3



However, psychometrically for the NCSC tests, the items do not exhibit a uniform increase in difficulty over tiers. Although there is a general pattern of items in lower tiers being easier than items in higher tiers, the relationships are irregular within content strands, and certainly inconsistent across strands. For example, as shown in Figure 2, there are reversals within content strands (e.g., Reading Literature tier 3 items have an average p-value of .56, but tier 4 items are easier on average, with an average p-value of .63. More often, there are not differences between tiers, such as between tiers 2 and 3 in Reading Informational Text or Foundational-Reading at Word Level. Notably, the tiers are not similar in difficulty across different content areas (e.g., tier 2 ranges from .46 in Foundational-Reading at Word Level to .78 for Writing Informational). To use tiers as organizing specifications may require a combination of more conceptual work and more careful development.

Figure 2: Average item difficulty by tier, ELA, Grade 3



Practical Requirements for Test Development

About 1000 student responses were required to calibrate an item and put it on the scale for NCSC 2015, using a two-parameter IRT model. It was possible to get that number of students for each item in the four forms because NCSC 2015 had so many states participate that over 4000 total students participated in each grade. Of course, the more forms with matrixed field-test items, the more items it is possible to try out and develop. Fewer potential students would mean fewer items could be developed with adequate data. For example, if a testing program had access to 1000 students per grade, and each form had slots for field-testing 5 items, then the testing program would be able to develop 5 new items per year at most; fewer, if not all the items survived field-testing. If the operational test had 30 items, then it would take at least six years for the testing program to develop enough items to completely replace a test form—if the exact

right items were developed and survived field-testing. Clearly, states with smaller numbers of students would be challenged to develop new items.

"Cousins," Clear Terminology, and Cooperation

Because NCSC has opted to allow users great flexibility in how they may use the NCSC 2015 testing materials, it is likely that different users will develop different "cousin" tests over time—some more closely related in terms of test blueprints, test characteristics, and implementation. Indeed, the NCSC 2015 tests re-administered by several states in 2016 are a "cousin" to the tests developed and administered by MSAA.

The states using assessment materials developed by NCSC should be careful in precisely describing or naming their assessments so that different "cousin" assessments are not all referred to as "a NCSC test" or (even worse) "the NCSC test." As "cousins" or new test forms are developed, it would be useful to have the developer/user describe how the new test is related to the NCSC 2015 assessment standards, specifications, and scale—that is, what is the degree of comparability intended between the two tests. It would be helpful to have empirical studies of relationships if possible.

Going into the future, users of post-NCSC test materials may strongly consider coordinating. They could coordinate and communicate to be sure there is clear terminology and communication about "cousin" variants, especially if they are not strictly comparable. Users may also wish to coordinate pooling of data if developing new items to allow more items to be developed more efficiently.