Evaluating the Quality and Impact of Items, Products, and Procedures:

NCSC Writing Alternate Assessment based on Alternate Achievement Standards[1]

Lori Nebelsick-Gullett, Elizabeth Towles-Reeves, Aminah Perkins, and Lauren Deters

edCount, LLC

A paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 2015.

The instruments referred to in this paper will be available as part of next steps in preparing the project technical documentation and for peer review. Anyone with an interest in these instruments at this time may contact the lead author at: lnebelsick-gullett@edCount.com.

[1] Disclaimer: The contents of this paper were developed as part of the National Center and State Collaborative under a grant from the U.S. Department of Education (PR/Award # H373X100002), Project Officer, Susan.Weigert@ed.gov. However, the content 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.

Abstract

NCSC used a research-to-practice model to support evidenced-based decision making. This paper describes the application of this model at key points in the development, revision, and refinement of the NCSC writing framework to ensure quality processes and products by understanding users' perspectives in addition to student results. The NCSC Alternate Assessment based on Alternate Achievement Standards writing framework addresses accessibility via administration procedures flexible enough to enable students to demonstrate what they know and can do. Several modes of inquiry were used to evaluate the quality of the processes and products. This paper highlights the following modes of inquiry: (a) writing task template tryouts, (b) writing evaluation study, and (c) a broader pilot of the writing items. These events are described and the outcomes are presented, followed by descriptions of how key findings were implemented and evaluated during subsequent studies. The overall impact on the design of the operational field test is also described.

Evaluating the Quality and Impact of Items, Products, and Procedures: NCSC Writing AA-AAS

The National Center and State Collaborative (NCSC) is a project led by five centers and 24 states[2] to build an alternate assessment based on alternate achievement standards (AA-AAS) for students with the most significant cognitive disabilities (SWSCD). NCSC's long-term goal is to ensure that SWSCD achieve increasingly higher academic outcomes and leave high school ready to participate in college, careers, and community. A well-designed summative assessment alone is insufficient to achieve that goal; an AA-AAS system also requires curricular and instructional frameworks as well as teacher resources and professional development. All partners share a commitment to the research-to-practice focus of the project and the development of a comprehensive model of curriculum, instruction, assessment, and supportive professional development. Partners designed the project to benefit participating states by yielding the first fully coordinated system of formative and summative assessments with curriculum, instruction, and professional development supports for improving achievement and outcomes for SWSCD. The partner states guided and shaped the work of the project in building this system, contributing to the effective implementation of research-to-practice activities, resulting in a model demonstration of innovative best practices.

Accessibility in the NCSC AA-AAS Framework

NCSC partners developed the NCSC AA-AAS writing framework with accessibility as a central goal throughout test design, development, and administration procedures. Accessibility allows a range of learners, with varying learner characteristics, to show what they know and to demonstrate their skills and abilities. By incorporating a principled design approach, NCSC integrated key principles that support accessibility such as, variable features and UDL. Variable features are aspects of assessment situations varied in order to control difficulty and/or to target emphasis on numerous aspects of knowledge, skills, and abilities required of students on the assessments. Incorporation of UDL principles ensures students have multiple means of representation, action, expression, and engagement.

NCSC completed reviews of extant literature, national content standards, and best practices to inform establishing measurement targets given reasonable opportunity to learn, however, additional modes of inquiry were required to further NCSC partners' understanding of student needs within this population, specifically their needs regarding accessibility (NCSC, 2015a). Ultimately, NCSC based the assessments on a research-to-practice model in the context of evidenced-based decision making. The purpose of this paper is to describe the application of the research-to-practice model at key points in the development, revision, and refinement of the NCSC writing framework. NCSC focused on developing a "working" definition of writing that

[2]The five NCSC partner organizations include: National Center on Educational Outcomes (NCEO) at the University of Minnesota, National Center for the Improvement of Educational Assessment (Center for Assessment), University of North Carolina at Charlotte, University of Kentucky, and edCount, LLC. The NCSC states participating in the Spring 2015 NCSC operational assessment are: Arizona, Arkansas, Connecticut, District of Columbia, Idaho, Indiana, Pacific Assessment Consortium, Maine, Montana, New Mexico, Rhode Island, South Carolina, South Dakota, and US Virgin Islands. As of Spring 2015, additional states are members of the NCSC Consortium, representing varying levels of participation. They are: California, Delaware, Florida, Louisiana, Maryland, New York, Oregon, Pennsylvania, Tennessee, and Wyoming.

would reflect an appropriate expectation of writing instruction throughout a student's educational experience and would be adaptable to the way in which SWSCD may produce writing. NCSC defines writing as generating a permanent product to represent and/or organize ideas or thoughts so messages are interpretable by someone else when the writer is not present. NCSC allows the use of symbols (e.g., picture symbols, objects) and assistive technology that produce text as writing. NCSC collected evidence to support evaluation of the processes and products of the writing framework to ensure quality by understanding both users' perspectives and student performance.

Development of Writing Core Content Connectors

NCSC developed Core Content Connectors (CCCs) that illustrate the core academic content knowledge defined by both the learning progression framework (LPF) and the Common Core State Standards (CCSS). NCSC based LPFs on research that describes how an understanding of core concepts in English language arts (ELA) and mathematics typically develops over time when students have the benefit of high quality instruction (Hess, 2010). Project partners prioritized up to 10 CCCs per grade and content area in grades 3-8 and high school for development of the NCSC AA-AAS. The prioritized CCCs, which link the CCSS and the model of domain-specific knowledge acquisition that guides academic instruction for these students, serve as the proximal assessment targets for the operational test in 2014-15. State partners used these CCCs as a starting point for designing the NCSC AA-AAS. For writing, NCSC partners prioritized three CCCs at each grade 3-8 and 11 (NCSC, 2015d) for assessment.

NCSC researchers used the Links for Academic Learning (LAL) model developed by Flowers, Wakeman, Browder, and Karvonen (2007) to conduct evaluations of the relationship between the Writing CCCs and the ELA CCSS. Overall, researchers found a strong relationship between the CCCs and CCSS. Because an individual CCC is often narrower in scope than an individual CCSS, researchers also found that an individual CCC sometimes exhibited a slight reduction in cognitive complexity when compared to its matched CCSS. Researchers recommended in some cases making small revisions to the CCC language to clarify the CCC or to address dual connections found in some cases.

Project developers used the results of the writing relationship study to support claims that the CCCs were targeted appropriately to the CCSS and in some cases developers refined CCC language based on researcher recommendations that arose from the evaluations. In addition, item developers used the results of this study to inform item creation. For example, developers ensured items aligned to the appropriate performance level and cognitive demands (depth of knowledge) of the CCC and CCSS.

Evaluating the Quality and Impact of the NCSC Writing AA-AAS

NCSC used several modes of inquiry to evaluate the quality and impact of the NCSC system in preparation for the operational test in spring 2015. Outlined in this section are key sources of data from studies NCSC used to inform the development and implementation of the writing portion of the NCSC AA-AAS: (a) task template tryouts, (b) writing evaluation study, and (c) a broader pilot of the writing items. This section describes the purpose, implementation, results, and impact of each study.

Writing Task Template Tryout

Using a principled approach to design based on evidence-centered design (ECD) literature (for example, Ewing, Packman, Hamen, & Thurber, 2010), NCSC developed Design Patterns and Task Templates that serve as item specifications. The task templates represent a "family" of items (i.e., items that assessed the same focal knowledge, skill, and ability) of various complexity levels. NCSC examined how students and teachers interacted with writing task templates (i.e., model items) developed for SWSCD in grades 3-8 and 11 (NCSC, 2015e). Within each task template, NCSC developed model task families as part of the initial phase of the AA-AAS development. For writing, NCSC developed both selected-response and constructed-response task families. The task families represented a gradation of complexity for assessing student knowledge, skills, and abilities in each measured standard. Developers created items at four levels of graduated complexity within each task template, using variable item features to ensure items within a template were broadly accessible for SWSCD. The purpose of the Writing Task Template Tryout Study was to identify and evaluate the content, complexity, and usability of items based on teacher review and administration of items to students. In fall 2013, NCSC conducted tryouts of the writing task templates developed using the principled design approach to address research questions in three areas: 1) Item feedback, 2) Student interaction, and 3) Teacher administration (See Table 1).

Researchers recruited 29 teacher volunteers from Arizona and South Dakota because these state partners represented different current practices in regard to alternate assessment. At the time of the study, Arizona did not have a writing assessment as part of their state alternate assessment. Conversely, South Dakota administered a formative writing assessment in grades 5, 7, and 10 for students taking the AA-AAS. Researchers asked each teacher volunteer to recruit three students representative of low, medium, and high disability levels, using their own criteria to determine these levels. (These criteria may have differed from teacher to teacher.) Teachers recruited 61 students, fewer students than the study's goals, due to classroom makeup in each of the states and the timing of the study.

After collecting informed consents, researchers mailed teachers the study materials and provided teachers with a 40-minute webinar training to describe the materials and administration process as well as answer teachers' questions. Teachers could begin to administer the items once they completed the training.

At the time of the study, developers had not created more than two task templates per grade so teachers administered only two items to each of their students. All but one student at each grade received a selected-response item and a constructed-response item.

Teachers administered the passages and items to their students over the course of two weeks. During item administration, teachers made note of student characteristics via the Learner Characteristics Inventory (LCI) and recorded student results via a Student Results Form. Student results included correct/incorrect responses, reactions to the items and passages, and general feedback on the items. Teachers returned materials to researchers once administration was complete.

Following item administration, researchers held one-hour teacher focus groups organized by state and grade span (grades 3-5 and 6-8 plus 11) via webinar. Focus groups included three types

of observers: a facilitator, a content expert, and a research expert. Finally, using a qualitative method to analyze the set of research questions, researchers examined the results from the item administration, compared the three types of focus group observers' notes, and created a summary.

Table 1: Writing Task Template Data Sources and Analyses by Research Question

Research Questions	Data Sources and Analyses
1. Do the items represent the full range of complexity needed to assess students' knowledge and skills?	• Student Results: Researchers examined the range of correct and incorrect responses by student ability level and passage level and noted any irregularities. Researchers also examined teacher notes taken during administration. • Teacher Focus Groups: Teachers explained item administration results. Researchers used teachers' focus group answers to verify teachers' Student Results Form comments.
2. Do the item directives provide enough information for administering the items?	• Student Results: Researchers organized, by item, notes that teachers took during item administration, especially those notes involving prompting or clarification. • Teacher Focus Groups: Researchers used narrative to present focus group comments about items.
3. Can various student response modes be accounted for by the writing prompts?	• Teacher Focus Groups: Researchers included major themes in narrative summaries.
4. Are the scoring rubrics appropriate for the student's ability level?	• Student Results: Researchers examined teachers' Student Results Form notes for comments about usability. • Teacher Focus Groups: Researchers used narrative to summarize teachers' comments about student-item interaction and related these comments to teachers' Student Results Form comments.
5. What suggestions do teachers have for improving the items?	• Student Results: Researchers examined and summarized teachers' Student Results Form notes. • Teacher Focus Groups: Researchers used narrative to summarize teachers' focus group comments.
Student Interaction	
6. Are there usability or accessibility issues that affect how students interact with the items?	• Student Results: Researchers integrated teachers' Student Results Form comments into the narrative that summarized teachers' focus groups comments. • Teacher Focus Groups: Researchers used narrative to summarize teachers' focus group comments.
7. Do the items allow the student to	• Student Results: Researchers compared

Research Questions	Data Sources and Analyses
demonstrate their writing skills?	notes on prompting/clarifications during item administration with item template descriptions to examine whether the cognitive process used to answer the question matched the skills designers intended the item to assess. • Teacher Focus Groups: Researchers used additional teacher input to corroborate issues found during administration.
8. Do the drafting templates effectively support a student's ability to draft a permanent writing product?	• Student Results: Researchers used narrative to summarize teachers' comments about students' interactions with items during administration. • Teacher Focus Groups: Researchers used narrative to summarize teachers' comments about student interactions.
9. Are usability or accessibility issues (if any) a potential source of construct-irrelevant variance?	• Student Results: Researchers used narrative to summarize teachers' Student Results Form comments. • Teacher Focus Groups: Researchers used teachers' focus group comments to corroborate their item administration comments.
Teacher Administration	
10. Are the variable features clear for administration?	• Teacher Focus Groups: Researchers used narrative to summarize teachers' focus group comments.
11. Where did teachers struggle the most in understanding how to present the item?	• Student Results: Researchers compiled and summarized teachers' item administration notes. • Teacher Focus Groups: Researchers linked teachers' focus group comments to their Student Results Form comments to verify consistency.
12. What suggestions do teachers have for adding other accommodations or adaptations?	• Teacher Focus Groups: Researchers used narrative to summarize focus group participants' suggestions.

Findings

NCSC identified the following findings based on the study:

- Almost all teachers agreed that the selected response (SR) items were easier for their students than the constructed (CR) items. One teacher commented that CR items were "spot on" on

for her moderate ability students, but that those same students giggled when presented with the easier SR items.

- The majority of teachers agreed that the items represented a range of complexity within the template. However, a few teachers indicated that the levels of complexity (referred to as tiers, in SR and CR do not correspond. Specifically, teachers indicated that the difficulty of a tier 2 SR item was not comparable to the difficulty of a tier 2 CR item.
- In general, teachers found the item directives thorough and helpful, albeit lengthy.
- The majority of teachers commented that they were unclear on how to appropriately use the rubrics, and that they could have benefitted from additional information regarding full and partial credit scoring.
- Most teachers reported that students enjoyed graphics and found them engaging, especially those in color.
- A few teachers expressed concern regarding administering items to students with unique needs. Specifically, one teacher commented that she had difficulty administering a CR item to her student who communicates via eye gaze, while another teacher questioned how she would administer CR items to students who use Augmentative and Alternative Communication (AAC).
- Teachers expressed uncertainty regarding whether the items allowed the students to demonstrate their writing skills. One teacher commented that the CR item gave him a sense of his student's writing skills, while another teacher indicated that the CR item allowed her students to demonstrate their organizational ability rather than writing ability.
- Teacher suggestions included: the addition of more pictures, the continued use of color pictures, providing an option to choose color or black and white pictures, allowing students to use a graphic organizer, providing additional space to draft a constructed response, and adapting the computer test to accommodate students who communicate via eye gaze.

Based on the results of the Writing Task template Tryout Study, researchers had the following recommendations for vendor partners and item writers:

- Consider minimizing the number of materials a teacher receives. Teachers found the testing materials and directions helpful, albeit lengthy.
- Investigate the complexity of the tiers for SR relative to the CR items. Even though teachers indicated that the items represented a range of complexity, many teachers remarked that the SR items were easier than the CR items.
- Consider flexibility in regards to graphics. Teachers overwhelmingly preferred color pictures, but a small minority asked for the option to choose black and white or color. Permit teachers to choose the pictures that allow their students to best access the items.
- As many teachers expressed confusion in using the rubric, clarify how to use the rubric. Consider additional professional development training or state-offered training regarding use of the rubric.

- Ensure students with alternative methods of accessing text and the items are able to access the items. Consider building in additional flexibility or accommodations – especially around the CR items - for those students who use eye gaze to communicate or an AAC device.
- Permit teachers to use more supports. One teacher requested the use of graphic organizers, while another teacher requested that her student use sequencing. Test developers should consider the variety of supports a student may use during administration.

Ultimately, researchers found that results from the writing task template tryout study confirmed that the items represented a range of complexity and that students engaged with the items, enabling them to demonstrate their writing skills. NCSC used the results to further develop administration processes, evidence capture, and designs for additional research.

Writing Evaluation Study

In the spring of 2014, NCSC designed a small-scale Writing Evaluation Study (WES) that was conducted as part of the initial pilot processes used for the NCSC AA-AAS development process (NCSC, 2015c). Researchers designed the study to include administration of both SR and CR writing items in multiple configurations of complexity to investigate students' interactions with and performance on the items. In the spring of 2014, NCSC administered a pilot of the ELA Reading test. A subset of students from 13 states who participated in the reading pilot also participated in the WES. Researchers identified six research questions displayed in Table 2 along with the data sources researchers used to address each question.

Table 2: Writing Evaluation Study Data Sources by Research Question

Research Questions	Data Sources
1. How does student performance compare across tiers of writing items?	Focus group responses, End of test survey
2. How do writing assessment task characteristics interact with student characteristics?	Focus group responses, Administration log
3. How well does student performance across the tiers align with teachers' *a priori* representation of student performance?	Focus group responses, Student response data, Administration log
4. How can the scoring process best recognize students' writing skills?	Focus group responses, Range finding
5. How are student writing scores related to selected response reading item scores?	Correlation of reading and writing performance scores
6. What lessons can we learn from the logistics of the writing task administration to enhance the operational administration?	Focus group questions, Range finding

Researchers used the student demographic data from the reading pilot and information on writing instruction to identify a group of possible teachers and students for participation in the study. Specifically, researchers designed the study so the student sample included students along the

continuum of current practices regarding writing instruction in the classroom. Researchers recruited 147 teacher volunteers to serve as test administrators (TAs) and 233 students for participation in the WES. Each TA was the classroom teacher during the testing year with up to two student participants.

Using items developed at the four tiers of complexity as previously described, researchers placed students in one of three groups (see Table 3) to organize and manage the distribution of items to students. Students were placed in groups based on specific student characteristics such as expressive and receptive language levels to ensure increased representation across the groups. At each grade level, the three group assignments were: Group 1) four less difficult CR items and four less difficult SR items, Group 2) four moderately difficult CR and four moderately difficult SR items, or Group 3) one CR and one SR item at each of the four difficulty levels. Researchers directed TAs to administer the writing items within a form in a specified order.

Table 3: Student Participation in WES

	Group			
Grade	**1**	**2**	**3**	**Total**
3	6	6	19	**31**
4	4	4	25	**33**
5	5	3	19	**27**
6	5	4	21	**30**
7	6	6	33	**45**
8	6	6	25	**37**
11	6	4	20	**30**
All	**38**	**33**	**162**	**233**

Researchers developed multiple instruments and processes to collect information regarding the WES Pilot test in addition to the actual items. The data sources included: (a) a Pre-administration Log, (b) an Administration Log, (c) an End of Test Survey (EOTS), (d) focus groups, and (e) input collected throughout range finding/scoring. Teachers used these instruments to provide in-depth feedback about student and test administrator experiences throughout the assessment administration process.

TAs completed a pre-administration log for each student prior to administering the items. Researchers developed the pre-administration log to contain questions asking TAs to predict student performance as well as questions focused on understanding student experience with writing instruction. Following administration of each item, TAs completed item specific questions in the Administration Log. Upon student completion of the assessment, TAs responded to the EOTS which included a series of questions regarding themselves as well as the student. After returning all study materials to researchers, TAs participated in focus groups to provide feedback on the test, training, student access to the test, and student participation in writing instruction during the school year.

For the final WES related activity, researchers conducted a range finding and scoring event to evaluate student performance on the items as well as the quality and appropriateness of the writing rubrics and scoring procedures. NCSC organizational and state partners along with experts with

experience in developing writing skills for SWSCD participated in this event. Each scorer independently scored a student product followed by a group discussion focused on addressing non-exact agreement; the discussion resulted in a consensus score. Researchers collected participant feedback through range finding and scoring. As a follow-up to the range finding and scoring event, participants met to discuss their observations and provide feedback regarding item directives as well as item content and structure.

Findings:
Researchers gathered the results from the various data sources associated with the WES and identified a plethora of information, for example:

- Across all grades, 203 TAs (92.3%) reported that they agreed or strongly agreed that they were confident in teaching the four assessed areas of writing.
- Across all grades, 73.2 percent of TAs either agreed or strongly agreed that their student actively engaged during classroom instruction of the assessed instructional areas in writing.
- Across all grades, approximately 80.0 percent of TAs reported that their students who participated in the WES use Standard English (alphabetic symbols) to create written products, as opposed to text in another language, Braille, pictures/symbols, selecting words from a list or chart, through means for dictation, or does not create written products.
- Across all grades, the most common assistive technology methods used by students during writing instruction included selection of pictures/symbols/objects from a list (31.4%) and selection of words from a list without pictorial representations (26.5%). TAs reported that 54.3 percent of their students participating in the WES do not use assistive technology to write (*n*=121).
- The majority of TAs (87.6% to 89.2%) indicated their agreement that the Directions for Test Administration (DTA) gave enough information for test administration. As a corollary, more TAs indicated that the DTA was helpful in preparing to administer the test than either the Test Administration Manual (TAM) or the training, though TAs also indicated that the TAM was helpful.
- The majority of TAs agreed that SR (>80%) and CR (>65%) items were of high quality.
- TAs in two of the three groups more often reported that their students were familiar with the selected-response item supports (82.5% to 83.8%) than with the constructed-response item supports (71.0% to 71.5%).
- TAs reported that the items represented an increase in complexity, noting that the more complex items had fewer supports.
- During the focus groups:
    - Over half of the participants indicated that their students found the selected-response items easier than the constructed-response items. Almost all TAs agreed that additional visuals/supports would assist their students in answering the items.
    - The majority of the participants indicated that they could score their student's work appropriately if provided with training and a detailed rubric that outlined the expectations for student writing.

  - When asked if their students would be able to access the writing items via an online assessment, the majority of the participants indicated that their students would be able to take the online assessment, depending on the accessibility accommodations.
  - Some participants shared that students had difficulty with understanding the purpose of revising and editing. They indicated that student fatigue occurred and may have impacted student performance.
- During range finding and the follow-up meeting:
  - Participants provided ongoing feedback to researchers that students writing products illustrated a range of knowledge, skills, and abilities.
  - Through discussion and revisions, content experts strengthened the alignment between the writing expectations and the scoring criteria to support consistent application of the rubrics to student products.
  - Participants proposed suggestions to improve the DTAs, student prompts, and stimulus materials.

Based on the results from the WES, researchers recommended the following next steps:

- Ensure a tight connection between the TA directives to students and the performance expectations delineated in the rubrics.
- Consider the processes and materials needed for TAs to score student work. Though TAs did not score student work, the majority of focus group participants indicated that they would be comfortable scoring their students' work using a rubric if they had examples of each level of work, time to score, and appropriate training.
- Consider using additional visuals for items, or allowing additional accommodations, which permit the use of more visuals. TAs reported the selection of pictures/objects/words from a list as the assistive technology students most used, and during focus groups, participants indicated that more visuals would help their students.
- As TAs most often reported the DTA and then the TAM as helpful in preparing for test administration, ensure that both the DTA and TAM provide clarity around specific item directions and permissible accommodations.
- Ensure that the online assessment provides access to all students with specific accessibility challenges.
- Make appropriate item-specific revisions to items, item directives, response templates, and stimulus materials based on discussions from range finding.

Project developers used results from the WES for evaluation and development purposes including refining the DTA and TAM, refining the rubrics for scoring writing items, and revising items. Project developers also used the recommendations that resulted from the WES to inform changes to the design of the pilot test in writing.

Broader Pilot of Writing Items

NCSC conducted a broader pilot of the writing items as part of the second round of NCSC AA-AAS pilot testing in the fall of 2014. The pilot test provided the final opportunity to gather evidence to inform the NCSC operational field test in spring 2015. Researchers designed the writing field test to carry forward the investigation of student performance across items of graduated complexity, to understand the relationship between reading and writing for students with the most significant cognitive disabilities, and to solidify the design of writing and scoring processes for the operational test in spring of 2015. Following the pilot test, TAs responded to survey questions specific to the writing test.

NCSC piloted writing items in two ELA sessions for all grades. Session One was comprised of reading items which were dichotomously scored, along with a set of writing selected-response items. The writing selected-response items were comprised of individual items scored in the NCSC assessment system (right vs wrong) as well as a set of selected-response items scored collectively to evaluate a student's ability to create a product using a series of connected questions. The sets of connected selected-response items were the tier 1 task intended to measure a student's ability to produce a written product. Session Two was comprised of a tier 2 constructed-response item and a tier 3 constructed-response item. Professional scoring staff scored these constructed response items (from tiers 2 and 3) using three trait scores: Organization, Idea Development, and Conventions. Possible scores for each of the three traits via a qualitative rubric included full evidence (F), partial evidence (P), limited evidence (L), or unrelated evidence (U). Researchers used a rubric to convert the patterns of performance across traits and tiers to a 0-7 score range.

Data Collection – Post Pilot 2 Scoring-related Activities

Following the pilot test, NCSC and the vendor developed a scoring plan describing the procedures for pre-range finding, range finding, and scoring of student work. Members of the NCSC content team reviewed several hundred student papers to identify a subset of papers for training and orienting the range finding committee members. Pre-range finding resulted in a set of student papers the vendor used with range finding participants to define the application of the scoring rubrics and to provide participants with an understanding of the tiers and grade-level expectations.

NCSC state partners were invited to participate in a range finding meeting which allowed for both on-site and virtual participation. The vendor's hand-scoring supervisors and the NCSC content leaders facilitated grade-level panel discussions during range finding. The objective of range finding was to identify sets of materials ―anchors, training, qualification and validity sets― to use during reader training and throughout the scoring event. The sets of materials helped ensure that the vendor's reader training and scoring were consistent with NCSC standards and guidelines.

The vendor's hand-scoring supervisors conducted reader training and qualification immediately preceding scoring for each item within a supervisor-led group. As a result, training was ongoing throughout the scoring period. All readers were required to qualify before scoring student responses for any item. The recommended qualification standard was 80% exact agreement with the predetermined score.

Readers scored all three rubric traits (organization, idea development, and conventions) during the same read. Two readers independently evaluated a given student response and a NCSC content lead served as the third reader when the trait scores for Reader 1 and Reader 2 were not in exact agreement. Researchers considered this resolution read the final read, resulting in the recorded score.

Findings – Selected Response

At each grade level, there were four selected response items, one at each tier. Test developers selected each item at a grade from a unique item family to ensure independence of items and prevent clueing. Developers selected the four items to represent the two CCCs measured by these item families. Items at tiers 1 and 3 were selected to address one CCC and items at tiers 2 and 4 addressed the other CCC. In investigating the writing selected response items at each grade, researchers found that overall, the *p*-values reflected the gradation of complexity across the tiers; that is, students tended to answer the lower-tiered items correctly more often than the higher-tiered items (refer to Table 4).

Table 4: Summary of P-Value by Tier of Selected-Response Writing Independent Items

Grade	Tier	Number of Items	P-Value
3	1	1	0.89
	2	1	0.75
	3	1	0.59
	4	1	0.58
4	1	1	0.68
	2	1	0.50
	3	1	0.57
	4	1	0.46
5	1	1	0.85
	2	1	0.69
	3	1	0.39
	4	1	0.47
6	1	1	0.70
	2	1	0.47
	3	1	0.64
	4	1	0.53
7	1	1	0.81
	2	1	0.61
	3	1	0.23
	4	1	0.50
8	1	1	0.86

Grade	Tier	Number of Items	P-Value
	2	1	0.54
	3	1	0.64
	4	1	0.45
11	1	1	0.71
	2	1	0.73
	3	1	0.30
	4	1	0.48

*Gradation in complexity was evaluated by comparing difficulty values within CCCs – tier 1 to tier 3 and tier 2 to tier 4.

Findings Constructed-response

Using the 0-7 score range, researchers calculated the average score points for student performance at each tier (see Table 5). (Note that the highest possible score for a tier 1 item was 2, the highest possible score for a tier 2 item was 5, and the highest possible score for a tier 3 item was 7.) Results showed that the average performance across the tiers ranged from 1.5-2.7, with tier 2 mean scores being generally higher than mean scores at the other tiers. Given the nature of the tier 1 writing product tier-connected item sets, it is difficult to directly compare performance means for tier 1 with those from tier 2 or tier 3. A comparison of mean values for tiers 2 and 3 indicated that, in general, students were able to produce higher quality products within the structure of the tier 2 items. Developers created the tier 2 items to be less complex and to provide additional guidance and support when compared to the tier 3 items. Mean score results support researchers' *a priori* assumption that the supports provided through the tier 2 items would result in more complete, higher quality written products for many students.

To deepen the understanding of the mean score comparisons, researchers analyzed frequency distributions for scores at each tier. When looking across all grades, researchers found that while over half of students received a score of 2 (out of maximum score of 2) for the tier 1 item sets, performance at the higher score ranges was rare for both tier 2 and tier 3 items. Analysis of scored responses showed that for both tiers 2 and 3, most scores fell in the range of 0-2. Researchers and content experts agreed that this highlighted the need for increased emphasis on writing instruction and opportunity to learn for SWSCD.

Table 5: Writing Item Statistics

Grade	Tier	Summary Statistics					
		N	Mean Total Score	SD Total Score	Median Total Score	Min Total Score	Max Total Score
	TIER 1	350	1.79	0.46	2.0	0	2
	TIER 2	92	2.20	1.72	2.0	0	5
3	TIER 2	73	2.71	1.68	2.0	0	5
	TIER 3	126	1.98	1.57	2.0	0	7
	TIER 3	110	1.69	1.49	2.0	0	6
	TIER 1	405	1.56	0.54	2.0	0	2
	TIER 2	108	2.03	1.53	2.0	0	5
4	TIER 2	109	2.32	1.70	2.0	0	5
	TIER 3	136	1.49	1.50	1.0	0	6
	TIER 3	148	1.95	1.56	2.0	0	6
	TIER 1	293	1.57	0.52	2.0	0	2
	TIER 2	83	2.46	1.55	3.0	0	5
5	TIER 2	62	2.02	1.65	2.0	0	5
	TIER 3	87	2.00	1.64	2.0	0	6
	TIER 3	114	2.08	1.69	2.0	0	6
	TIER 1	281	1.71	0.51	2.0	0	2
	TIER 2	67	2.54	1.46	2.0	0	5
6	TIER 2	61	2.51	1.40	2.0	0	5
	TIER 3	95	1.86	1.39	2.0	0	7
	TIER 3	110	1.60	1.50	2.0	0	7
	TIER 1	319	1.81	0.42	2.0	0	2
	TIER 2	69	2.16	1.54	2.0	0	5
7	TIER 2	74	2.45	1.58	2.0	0	5
	TIER 3	128	1.63	1.37	2.0	0	7
	TIER 3	111	1.65	1.29	1.0	0	7
	TIER 1	422	1.88	0.35	2.0	0	2
	TIER 2	89	1.89	1.32	2.0	0	5
8	TIER 2	100	2.21	1.46	2.0	0	5
	TIER 3	141	1.84	1.69	1.0	0	7
	TIER 3	178	1.96	1.38	2.0	0	7
	TIER 1	378	1.88	0.35	2.0	0	2
	TIER 2	92	1.80	1.58	1.0	0	5
11	TIER 2	96	1.97	1.48	1.0	0	5
	TIER 3	132	1.72	1.70	1.0	0	7
	TIER 3	136	1.86	1.56	1.5	0	7

Quality and impact of items, products, and procedures

Findings from the ELA/Writing end of test survey include:

- When asked about writing instruction, 63.4 percent of TAs reported that their students received moderate to considerable focus on Conventions of Standard English across the grade spans (grades 3-8 and 11). However, only 32.0 percent of TAs reported their students received moderate to considerable focus in narrative/fiction writing, only 26.7 percent of TAs reported moderate to considerable focus in explanatory writing, and only 9.7 percent of TAs reported moderate to considerable focus in argument based writing.
- When asked who entered the constructed-response information into the online template, 40.8 percent of TAs reported that the student entered the information directly into the online template, 33.6 percent of TAs reported that they entered the information based on the student's oral response, and 15.7 percent of TAs uploaded their student's paper-based product.
- When asked if the student found the tier 1 constructed-response items easy or hard, 85.6 percent of TAs responded that the students found most or all of the items hard. When asked whether their students found the tier 2 and tier 3 items easy or hard, 85.9 percent of TAs reported that the students found most or all of the items hard.
- Slightly less than half (47.0 percent) of TAs agreed that their student was able to actively engage with the constructed-response items.

Based on the results from the pilot test and the corresponding end of test survey, researchers recommended the following next steps:

- Clarify the language in the items' test directives such that the expectations, as described in the rubric, are clear to the student.
- Strengthen the reader training protocol and provide more time for readers to gain familiarity with the scoring rubrics and an understanding of the tiers and grade-level expectations.
- Strengthen the rubrics for tiers 2 and 3 to ensure that the rubric descriptive statements provide clear performance expectations for students as well as a logical continuum.
- Continue to promote students' opportunity to learn. Based on survey responses, students receive less instruction in Narrative/Fiction, Explanatory, and Arguments; TA responses regarding difficulty of the items may be reflective of the lack of instructional focus.

Conclusions and Next Steps

Across the NCSC studies, staff gathered evidence regarding the appropriateness of the writing item content and complexity, the usability and accessibility of the items, the ease of item administration for both students and teachers, and the capacity of SWSCD to show achievement in writing.

The task template study demonstrated initial models for how to best assess writing, and the data provided information regarding (a) how well the items reflected differing levels of complexity; (b) improvements to the items, materials, and scoring rubrics; (c) student engagement and accessibility; and (d) administration fidelity. These results informed the WES study, leading

researchers to continue to investigate complexity across item families and to seek a deeper understanding of student performance. The WES study also provided an opportunity to evaluate the utility and clarity of the writing DTAs. As a result of the WES study, developers reviewed and revised all directives to enhance clarity, reevaluated accessibility options for presenting the writing test online, and identified revisions to the items and rubrics. The writing pilot test reflected the improvements made to the items, the administration directions and procedures, and supports for accessibility. Results from the writing pilot test indicated a need for NCSC to provide clear expectations and the definition of writing to teachers of students participating in the AA-AAS. In addition, researchers suggested NCSC revise the test administrator training to ensure TAs are prepared to manage the idiosyncrasies of the online platform and carefully evaluate the writing items against the data trends to ensure that readability and comprehensibility are maximized and scoring accuracy is optimized.

References

Cameto, R., Haertel, G., DeBarger, A. & Morrison, K. (2011). Alternate assessment design-mathematics technical report 1: Project overview applying evidence-centered design to alternate assessments in mathematics for students with significant cognitive disabilities. In *How design patterns integrate universal design for learning (UDL) into assessments for students with disabilities.* Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*(4), 325-341.

Flowers, C., Wakeman, S., Browder, D. & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, North Carolina: University of North Carolina at Charlotte.

Hess, K. K. (December 2010). *Learning progressions frameworks designed for use with the common core state standards in mathematics K-12.* National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment, Dover, N.H. (updated- v.2).

National Center and State Collaborative. (2015a). *Assessment accessibility annotated bibliography.* Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

National Center and State Collaborative. (2015b). *NCSC ELA pilot 2 end of test survey results.* Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

National Center and State Collaborative. (2015c). *NCSC pilot 1 spring 2014 writing evaluation study report.* Minneapolis, MN: University of Minnesota, National Center and State Collaborative

National Center and State Collaborative. (2015d). *Study of the relationship among the writing core content connectors and the English/Language Arts common core state standards.* Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

National Center and State Collaborative. (2015e). *Writing task template tryout study.* Minneapolis, MN: University of Minnesota, National Center and State Collaborative.