Technical Report 64

Rules for Audio Representation of Science Items on a Statewide Assessment: Results of a Comparative Study

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Supported by:

U.S. Office of Special Education Programs

Technical Report 64

Rules for Audio Representation of Science Items on a Statewide Assessment: Results of a Comparative Study

Christopher Johnstone • Christopher Rogers • Yi-Chen Wu •
Gaye Fedorchak • Michael Katz • Jennifer Higgins

June 2012

Johnstone, C., Rogers, C., Wu, Y.-C., Fedorchak, G., Katz, M., & Higgins, J. (2012). *Rules for audio representation of science items on a statewide assessment: Results of a comparative study* (Technical Report 64). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The contents of this report were developed under an Enhanced Assessment Grant from the U.S. Department of Education to the State of New Hampshire. The contents do not necessarily represent the policy of the U.S. Department of Education and you should not assume endorsement by the Federal government.

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through Cooperative Agreements (#H326G050007 and #H326G110002) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. This report was funded with partial support from the State of New Hampshire and Nimble Tools on a Grant for Enhanced Assessment Instruments funded by the U.S. Department of Education from January 2009–June 2010. Additional support for targeted projects, including those on English language learners, is provided by other federal and state agencies. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

IDEAs that Work
U.S. Office of Special Education Programs

National Center on Educational Outcomes
University of Minnesota • 207 Pattee Hall
150 Pillsbury Dr. SE • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://www.nceo.info

Executive Summary

Large-scale assessment practice has moved consistently from a paper-and-pencil exercise to online assessments over the past decade. New formats for testing allow for new opportunities to provide students with disabilities access to items so that they may most validly demonstrate their knowledge. In this study, we investigated an online auditory feature of an assessment, designed to provide students with challenges with print reading with content information. In an effort to evaluate the impact of how content is presented in auditory fashion, NimbleTools (now the Measured Progress Innovation Lab) and the National Center on Educational Outcomes examined three approaches to "scripting" or creating audio representations of items through a New Hampshire led Enhanced Assessment Grant. We found that students with disabilities' mean scores were .13 higher on items in which tables were read automatically than when they were not. Mean scores for students with disabilities were .21 higher for items in which chemical equations were read as letters than when items were read as words audio presentation, and mean scores were .05 higher when units of measures were read in tables than tables that did not have audio presentation. None of these findings were statistically significant, and student preference for auditory scripting only sometimes aligned with achievement (i.e., sometimes student preferences correlated with achievement results while other times students preferences correlated with lower comparable achievement).

Table of Contents

Overview

The exclusion of students with disabilities from large-scale assessments is a practice that has nearly disappeared over the past two decades. Assessment reforms from the 1990s to present have both encouraged and enforced the participation of students with disabilities in large-scale assessments. Despite the upswing and maintained high rate of participation of students with disabilities in large-scale assessments, these students still lag behind their peers without disabilities (Thurlow, Quenemoen, Altman, & Cuthbert, 2008; Thurlow, Bremer, & Albus, 2011).

Explanations about why students with disabilities score lower on state assessments than their peers without disabilities focus on a variety of aspects, but one particularly worrisome reason is because of factors within the test itself. Messick (1989, 1996) first noted that increasing achievement on tests should reflect increases in student knowledge of the subject matter. Likewise, decreases in scores should correlate with decreased knowledge of the subject matter, not a factor unrelated to the construct itself. Messick explained that construct-irrelevant error may be present when test items require skills that go beyond the intended construct. Such skills may produce challenges for certain students because of the characteristics of their disability.

In an effort to reduce construct-irrelevant variance for students with disabilities, all states now allow testing accommodations for students with demonstrated needs. The use of accommodations can be described as an attempt to ensure that the scores received by students with disabilities are valid measures of achievement (Christensen, Braam, Scullin, & Thurlow, 2011). For states that use paper-based tests, there are often two clear choices for test administration: standard and with accommodations. Thurlow and Bolt (2001) defined testing accommodations as:

> changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests. Accommodations attempt to eliminate barriers to meaningful testing, thereby allowing for the participation of students with disabilities in state and district assessments (p. 3).

Historically, testing accommodations have been addressed after-the-fact on tests, that is, tests were created first, followed by the implementation of accommodations. This is sometimes described as "retrofitting," a metaphor borrowed from architectural terminology to describe changes made to structures after they have been built to make them more accessible. In an effort to promote more accessible assessments from the beginning, scholars began in the early 2000s to use the term *universal design of assessment*. Universal design of assessment (UDA) is broadly defined as assessments that are "designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment" (Thompson, Johnstone, & Thurlow, 2002, p. 5).

As technology-based assessments became prevalent in state testing systems, Dolan, Hall, Banerjee, Chun, and Strangman (2005) theorized that universally designed assessments were those that allowed students access to the assessment constructs through multiple pathways of representation and response. Put simply, Dolan et al. believed that mode of presentation should not hinder a student's opportunity to access assessments. One way to provide access to students is through audio functions on computer-based tests. Although audio accommodation tools have been present for many years (Christensen et al., 2011), when audio functions on computer-based assessments are conceptualized as a way to "allow participation of the widest possible range of students" (Thompson et al., 2002), the accommodation becomes a standard feature of a universally designed assessment.

Test item accessibility is a requirement of the Race to the Top Assessment (RTTA) and the General Supervision Enhancement Grants (GSEG) that support consortia of states to develop assessments. The efforts of the RTTA and GSEG Consortia provide an unprecedented opportunity to positively impact the accessibility of assessment practices. Common rules for audio representation of science items do not currently exist. The closest example of rules for audio representation of items can be found in the National Assessment of Educational Progress 2009 *Manual for Assessment Administrators*. Guidelines are needed to ensure that auditory representation provides high-quality access to test content, leading to more reliable and valid assessments.

This conceptualization was the theoretical underpinning for the study reported here. NimbleTools is a test delivery system that was designed with universal design principles to provide access to content for all students through a variety of accessibility tools (including features that mask parts of the item, on-demand calculators, magnification, auditory calming, and audio versions of items). In an effort to rigorously evaluate and define rules for audio representation of items in a standardized way, NimbleTools and the National Center on Educational Outcomes examined three approaches to scripting or creating audio representations of items. Student achievement on items presented using different "scripting" rules was compared. Further, student perceptions on these rules were solicited.

Overview of Study

In June 2010, students from three schools in the northeast United States participated in a computer-based science practice assessment that allowed for audio versions of all test items for all students. Students completed the assessment by reading (visually) and listening to audio versions of test items. Items were presented using two or three rules for presentation. Students with and without disabilities participated in the study. All students received a tutorial on how to use the functions of the NimbleTools delivery system prior to taking the actual test.

Sample

A sample of 93 students participated in the study. All of these students attended public high school in the northeast U.S. and all were in grade 11. Four test forms were used in this study, and students were randomly assigned to each form. For comparative purposes, students were asked on a demographic survey whether they had an Individualized Education Program (IEP). In order to have an IEP, a student must receive some form of special education service. In total, 19 students (n=19) indicated that they had an IEP (and therefore, some form of disability). The majority of students (n=46) stated that they did not have an IEP. Many of the students (n=31) responded that they did not know whether they had an IEP. For purposes of analysis, we included these students in the group of students without disabilities. Our rationale for this decision is that the term IEP is ubiquitous in special education programming. A child in grade 11 will likely have begun receiving special education services long before grade 11 and would therefore be familiar with the term IEP from attending meetings, working on IEP goals with teachers, and signing IEP forms (required after age 14 in most states). Therefore, to our best understanding, the sample was n=74 (students without disabilities) and n=19 (students with disabilities).

Table 1 below provides demographic information on the students, including gender, ethnicity, whether the students have ever used audio functions, and whether they had used it recently, along with the form of the test students took (randomly assigned). A closer look at the sample in Table 1 reveals that the sample had slightly more females than males, was predominately White, and was inexperienced with audio tools at the time the research took place.

Table 1: Demographic Characteristics of Sample

Number of students by different background variables in each form

		Form 1	Form 2	Form 3	Form 4	Total	
		N	n	n	n	N	%
Gender	Female	22	6	4	18	50	53.8
	Male	11	7	8	17	43	46.2
Ethnicity	Asian	0	0	0	3	3	3.2
	Black or African American	0	0	0	2	2	2.2
	Hispanic	1	1	0	1	3	3.2
	Native Hawaiian or Pacific Islander	1	0	0	0	1	1.1
	White	29	11	12	29	81	87.1
	Others	2	1	0	0	3	3.2
IEP	I don't know	16	2	4	9	31	33.3
	No	12	8	4	19	43	46.2
	Yes	5	3	4	7	19	20.4

Table 1: Demographic Characteristics of Sample (continued)

Number of students by different background variables in each form

		Form 1	Form 2	Form 3	Form 4	Total	
		N	n	n	n	N	%
Read Aloud Ever	No	31	10	8	29	78	83.9
	Yes	2	3	4	5	14	15.1
	Missing	0	0	0	1	1	1.1
Read Aloud Recent	No	29	12	11	32	84	90.3
	Yes	4	1	1	3	9	9.7
	Total	33	13	12	35	93	100.0

Procedures

Definition of Scripting Rules

Students completed items using three types of skills: (a) solve a problem using a table, (b) solve a problem that involves chemical equations, and (c) solve a problem with units of measure found in a table. Each of these skills were found in state science standards for Grade 11 and were therefore included in the state assessment. For each of the skills assessed, a scripting rule was developed.

For example, when students were required to solve problems using tables, the reader (the person voicing the item) would either: (a) read table elements in the flow of how they are presented in the item, (b) not read elements in table, but allow students to click on specific elements as needed for audio representation, or (c) re-order items so that text is presented prior to the reading of the table. Likewise, if chemical equations were present in items, the reader would either (a) read the equation as a "word" (e.g., Na = sodium)[1]; or (b) read equations as letters (e.g. Na = "en-ay"). Finally, for tables with units of measure, units were either (a) read by a professional reader; or (b) not read. The text of all items was read in chunks, meaning that a human reader presented information in natural sentences and chunks of text (generally sentences or clauses). These "chunks" were highlighted on the screen simultaneously with audio presentation. Table 2 demonstrates the various scripting rules associated with each skill requirement. Appendix A contains specific examples of how scripting rules were implemented.

[1] In this case, scripting provided the audio representation of the chemical symbol represented. For the sake of this report, audio representations of symbols are called "words."

Table 2: Scripting Rules

	Scripting Rule 1	Scripting Rule 2	Scripting Rule 3
Tables/diagrams	Read table elements in the flow of how they are presented in the item.	Do not read table elements when they are presented, but allow students to click on individual elements to have them read when requested.	Re-order item to have introductory text and question presented prior to presentation of table. Then read table elements as they are presented.
Chemical equations	Read chemical equations in "words" instead of in symbols. For example, in an equation, read Na as "sodium" instead of "N-a"	Read chemical equations as letters.	
Units of measure in a table	Read units of measure associated with table elements.	Do not read units of measure associated with table elements.	

Source: Nimble Innovation Lab, Measured Progress

Instrumentation and Methods

As noted, all students completed a practice science assessment using the computer based delivery system NimbleTools. This assessment mimicked the statewide assessment in content and level of difficulty, and the NimbleTools system allowed all students to use the audio function on the test. Data were collected electronically using the NimbleTools interface. Students answered questions online, and answers were uploaded into a database. At the conclusion of each section, all students completed two survey items pertaining to student preferences for scripting rules.

Students were randomly assigned to one of four test forms. Each form contained the same items, in the same order, but with different variations of scripting rules. Students were asked to complete nine test items and two survey items focusing on audio representation of tables, six test items and two survey items focusing on audio representation of chemical equations, and six test items and two survey items focusing on audio representation of units of measure for table elements. In total, students completed a 27-item test form. After these 27 items, there were four questions to gather students' demographic information.

As seen in Table 3, there were three versions scripted under *Tables/diagrams*: (a) automatically read, (b) do not automatically read, and (c) restructure item to present graphic/table after prompt. Therefore, 1A = item 1 scripted to be read automatically, 1B = item 1 scripted to be read not automatically, and 1C = item restructured. Under the element of *Chemical equations*, there were two scripted versions—words vs. letters, and two scripting versions of *Units of measure in a table*—units of measure vs. no units of measure.

Table 3. The Structure of Each Test Form

Content elements	Item number	Form 1	Form 2	Form 3	Form 4
Tables/Diagrams	Item 1	1A	1B	1B	1C
	Item 2	2A	2B	2B	2C
	Item 3	3A	3B	3B	3C
	Item 4	4B	4C	4C	4A
	Item 5	5B	5C	5C	5A
	Item 6	6B	6C	6C	6A
	Item 7	7C	7A	7A	7B
	Item 8	8C	8A	8A	8B
	Item 9	9C	9A	9A	9B
Survey items		The nine items that you completed read aloud items in two different ways. For some items the details of the tables and diagrams were automatically read aloud to you, for other items the details were not automatically read aloud to you.			
	Item 10	Which way do you prefer[a]? (options “automatically read”, “not automatically read”, or “no preference”)			
	Item 11	If you had to choose one of these ways to have items read to you on a test, which way would you choose? (options “automatically read”or “not automatically read”)			
Chemical Equations	Item 12	12A	12A	12B	12B
	Item 13	13A	13A	13B	13B
	Item 14	14A	14A	14B	14B
	Item 15	15B	15B	15A	15A
	Item 16	16B	16B	16A	16A
	Item 17	17B	17B	17A	17A
	Survey items	The six items that you just completed read aloud chemistry equations in two different ways. For some items, the letters representing the chemistry symbols were read to you, for other items the words representing the chemistry symbols were read to you.			
	Item 18	Which way do you prefer? (options “letters representing chemistry symbols”, “words representing chemistry symbols”, or “no preference”)			
	Item 19	If you had to choose one of these ways to have items read to you on a test, which way would you choose? (options “letters representing chemistry symbols” or “words representing chemistry symbols”)			
Chemical Symbols	Item 20	20A	20A	20B	20B
	Item 21	21A	21A	21B	21B
	Item 22	22A	22A	22B	22B
	Item 23	23B	23B	23A	23A
	Item 24	24B	24B	24A	24A
Chemical Symbols (continued)	Item 25	25B	25B	25A	25A
	Survey items	The six items that you just completed read aloud table elements in two different ways. For some items, units of measure were read to you with each table element, for other items units of measure were not read to you with each table element.			
	Item 20	Which way do you prefer? (options “read units of measure with table elements”, “do not read units of measure with table elements”, or “no preference”)			
	Item 21	If you had to choose one of these ways to have items read to you on a test, which way would you choose? (options “read units of measure with table elements” or “do not read units of measure with table elements”)			
Demographic Questions	Gender (male, female) Ethnicity (American Indian or Alaskan Native, Asian, Black or African American, Hispanic, Native Hawaiian or Pacific Islander, White, Other) Do you have an IEP? (yes, no, I don’t know) Have you ever used a read aloud accommodation during testing? (yes, no) Did you use a read aloud accommodation for any 2009-2010 (state) test? (yes, no)				

[a] Because students would not have known that the item was restructured and this idea would have been difficult to explain. All restructured items were scripted “not automatically read,” which is how students would have identified them.

A small sample of students (n=16) was administered a truncated version of the test, which had just nine items. The nine item form presented one or two items in each scripting category. The students interviewed were not included in statistical analyses, but were interviewed while they took the mini-assessment, using cognitive lab methods. Using these methods the students worked through each test item and were asked to say everything that came to their mind. Ericsson and Simon (1994) noted that asking students to comment while cognitive activities are still in their short-term memory is valuable for getting honest feedback on stimuli, because as soon as perceptions reach longer term memory they are tainted by other perceptions.

Once students completed items, they were asked a series of follow-up questions. Almond, Cameto, Johnstone, Laitusis, Lazarus, Nagle, Parker, Roach, and Sato (2009) noted that having a follow up interview also provides valuable information because think aloud data are easier to understand and interpret. Retrospective questions produced less spontaneous data than utterances produced while students were solving problems, but helped us to fill in any missing gaps in data from the “live” data produced while students answered items.

Analysis

Quantitative Analysis

A two-group design was employed to estimate the effect that the application of a given scripting rule had on the item difficulty for each item. Analyses focused on the effect that different scripting rules had on student achievement and student preferences. The main research question that guided quantitative analysis was "Is there a difference in student preference regarding how items are presented in auditory form?" Under this main question, there were several sub-questions:

1. Is there a main effect of disability on student preference for each content element (*Table/diagram, Chemical equations*, and *Units of measure in a table*)?
2. Is there a main performance effect of how items are presented in auditory form on student preference for each content element (*Table/diagram, Chemical equations*, and *Units of measure in a table*)?
3. Is there a performance interaction effect of disability status and how items are presented in auditory form on student preference for each content element (*Table/diagram, Chemical equations*, and *Units of measure in a table*)?

A descriptive analysis was performed to see the trend in student performance among different scripting rules in each content element (Tables/diagrams, Chemical questions, and Chemical symbols). Because students were tested on different scripting rules within each form, a repeated measures approach was used to analyze the data. The repeated measures were performed to test the difference in performance for different scripting rules in each content element.

Further, a descriptive analysis was performed on the survey items to calculate the percentage of each response in each survey question. Chi-square analyses were also performed to test whether the response patterns differed by disability status.

Qualitative Analysis

A point-by-point analysis was conducted on the qualitative data. In this analysis, student utterances were analyzed one by one for content. Those utterances were then coded according to student preference, student observations about scripting approaches, and other student comments. Student performance was not examined during the qualitative portion of this study. Rather, interviews were used to specifically understand the cognitive responses that different scripting approaches elicited in students.

Results

Content Element

In all content areas, students without disabilities scored higher on every category of item than students with disabilities. Descriptive data also demonstrated that performance patterns differed by populations. For example, students without disabilities scored almost equally for all rules about Tables and Diagrams (M=1.78 for "read not automatically" compared to M=1.76 for other scripting rules). However, students with disabilities scored higher on Tables and Diagrams that were read automatically (M=1.68 compared to M=1.53 for other categories). Likewise, students without disabilities scored almost equally for chemical equations (M=1.47 for equations read as words compared to M=1.45 for equations read as letters), while students with disabilities performed better when chemical equations were read as letters (M=1.16 compared to M=.95). Finally, students without disabilities scored higher when units of measure were not read in tables (M=1.42 compared to M=1.34). This was a different pattern from that for students with disabilities, who scored higher when units of measure were read (M=1.21 to M=1.16 units of measure not read). Table 4 presents the sample size, mean, and standard deviation of student performance on items by content element and scripting rule. Figures 1-3 present mean scores by disability status and scripting rules in each content element.

Table 4. Descriptive Analysis of \Students' Performance by Content Element and Scripting Rule

		Non-SPED (n=74)		SPED (n=19)		Total (n=93)	
Content element	**Scripting rule**	M	SD	M	SD	M	SD
Tables/Diagrams	Read automatically	1.76	0.90	1.68	1.06	1.74	.93
	Read NOT automatically	1.78	0.91	1.53	0.84	1.73	.90
	Restructured	1.76	0.87	1.53	0.70	1.71	.84
	Total	5.30	1.99	4.74	2.00	5.18	1.99
Chemical equations	Letters	1.45	0.97	1.16	1.01	1.39	.98
	Symbols (words)	1.47	0.98	0.95	0.97	1.37	1.00
	Total	2.92	1.59	2.11	1.41	2.75	1.59
Units of Measure	Units of measures	1.34	0.90	1.21	0.98	1.31	.91
	No units of measures	1.42	1.05	1.16	0.76	1.37	1.00
	Total	2.76	1.54	2.37	1.16	2.68	1.48

Figure 1. Mean Scores of Scripting Rules for Tables/Diagrams and Student Disability Status

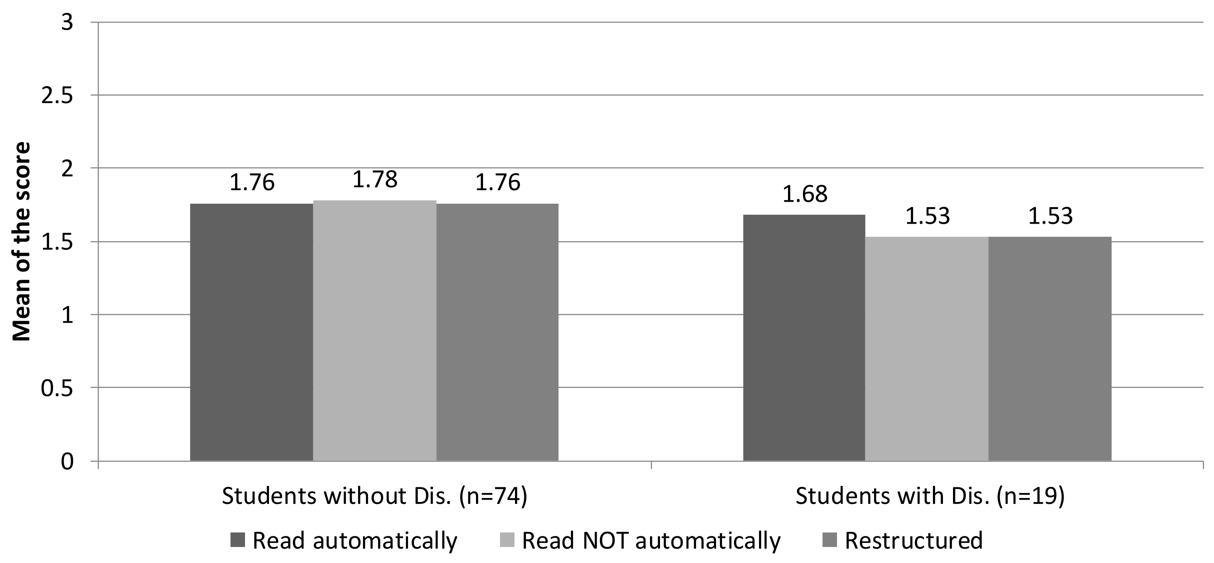


Figure 2. Mean Scores of Scripting Rules for Chemical Equations and Student Disability Status

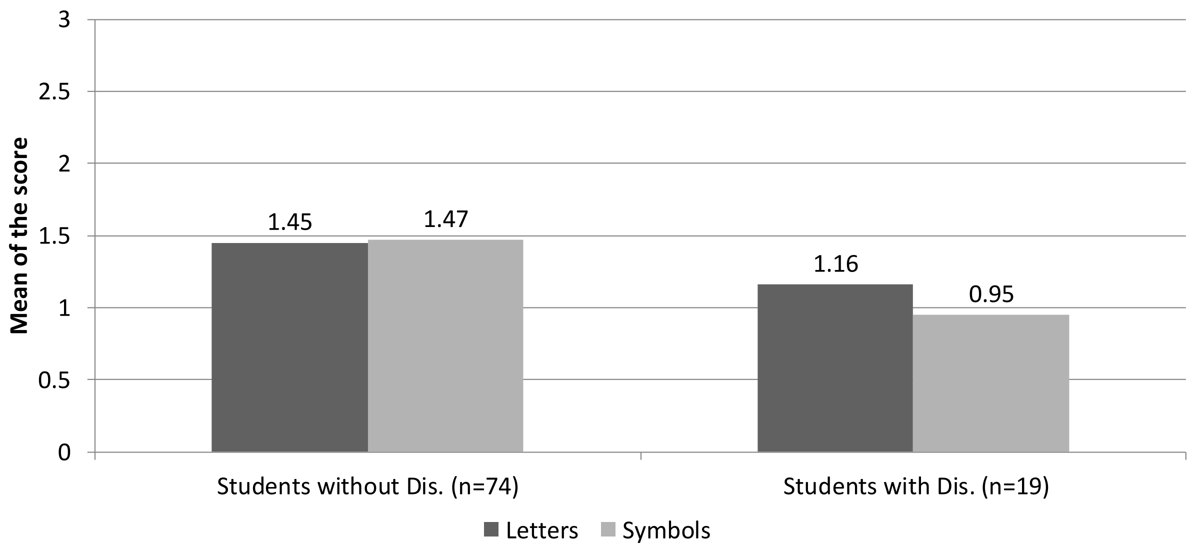
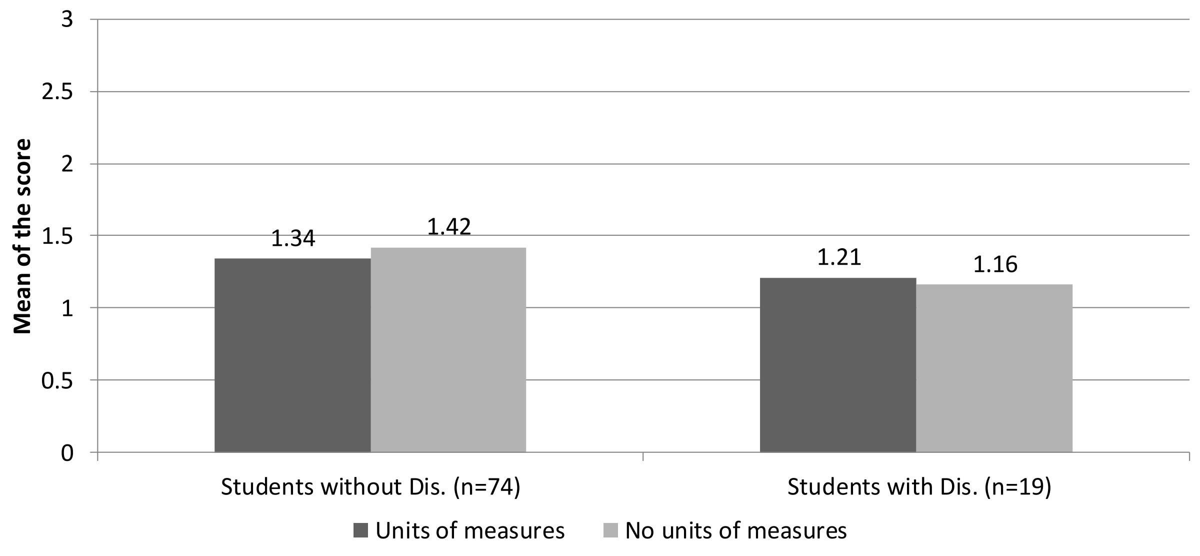


Figure 3. Mean Scores of Scripting Rules by Units of Measure and Disability Status



Repeated Measures Study

In order to test differential impact of scripting rules, repeated measures were used to determine whether differences in scores were statistically significant. Results indicated that there was no main effect of the scripting rules for *tables/diagrams* (read automatically, read not automatically, restructured; $F = .20$, $p > .05$, $df = 2$). Further, there was no interaction between disability status and scripting rules for tables/diagrams ($F = .28$, $p > .05$, $df = 2$). Finally, there was not a statistically significant difference in student results by disability status ($F = 1.20$, $p > .05$, $df = 1$).

There was no main effect for the scripting rules related to *chemical equations* (words vs. letters; $F = .36$, p > .05, $df = 1$). There was also no interaction effect between disability status and chemical equations scripting rules ($F = .611$, $p > .05$, $df = 1$). However, there was a main effect on disability status ($F = 4.12$, $p < .05$, $df = 1$). That means the students without disabilities statistically significantly outperform students with disabilities under all scripting conditions.

Finally, there was no main effect for the scripting rules about *units of measure* (units vs. no units; $F = .01$, p > .05, $df = 1$). There was also no interaction between the two variables of disabilities and reading/not reading units of measure ($F = .18$, $p > .05$, $df = 1$). Finally, there was not a statistically significant difference in achievement on relevant items for this section between students with and without disabilities ($F = 1.05$, $p > .05$, $df = 1$).

Results indicate that scripting rules did not have statistically significant impacts on student performance. However, in one category of items, students with disabilities performed better in "units of measures," but students without disabilities performed better on "no units of measures."

This distinction may provide insights into opportunity to learn, cognitive load while completing assessments, and support needs for students with disabilities.

Survey Questions

Chi-square results indicated that there were no statistically significant differences in preferences, by item, for scripting rules. Table 5 presents the percentage and number of students choosing particular scripting rules and relevant chi-square results. Chi-square tests were performed to test whether the percentages of preference between students with and without disabilities differed by item.

Due to the small cell size for students with disabilities, statistical error may be present. Although differences were not statistically significant between populations, there were descriptive trends in the data (which is presented in tabular form below). For example, for rule preference survey questions with tables/diagrams, most students preferred the details of the item to not be automatically read aloud (62.4%). Further, overall students preferred the words representing the chemistry symbols to be read to them in chemical equations (66.7%). There was only a very slight majority of students who preferred the units of measure to be read to them with each table element (50.5%). Preferences of students, disaggregated by student population and scripting preference, are presented in Figures 4-6.

Table 5. Percentage and Number of Student Responses to Survey and Chi-square Results

Survey Question	Options	Non-SPED		SPED		Total		
		N	%	n	%	N	%	χ^2
Survey items after "in flow" test items (1-9): The nine items that you completed read aloud items in two different ways. For some items the details of the tables and diagrams were automatically read aloud to you, for other items the details were not automatically read aloud to you.								
Which way do you prefer?	Automatically read	14	18.9	6	31.6	20	21.5	1.44$^{\text{n.s.}}$
	No preference	19	25.7	4	21.1	23	24.7	
	Not automatically read	41	55.4	9	47.4	50	53.8	
If you had to choose one of these ways to have items read to you on a test, which way would you choose?	Automatically read	27	36.5	8	42.1	35	37.6	.20$^{\text{n.s.}}$
	Not automatically read	47	63.5	11	57.9	58	62.4	
Survey items after "chemistry equation" test items (10-15) The six items that you just completed read aloud chemistry equations in two different ways. For some items, the letters representing the chemistry symbols were read to you, for other items the words representing the chemistry symbols were read to you.								
Which way do you prefer?	Letters representing chemistry symbols	15	20.3	4	21.1	19	20.4	.08$^{\text{n.s.}}$
	No preference	24	32.4	6	31.6	30	32.3	
	Words representing chemistry symbols	35	47.3	9	47.4	44	47.3	
If you had to choose one of these ways to have items read to you on a test, which way would you choose?	Letters representing chemistry symbols	25	33.8	6	31.6	31	33.3	.03
	Words representing chemistry symbols	49	66.2	13	68.4	62	66.7	
Survey items after "chemistry equation" test items (16-21) The six items that you just completed read aloud table elements in two different ways. For some items, units of measure were read to you with each table element, for other items units of measure were not read to you with each table element.								
Which way do you prefer?	Do not read units of measure	31	41.9	6	31.6	37	39.8	1.49$^{\text{n.s.}}$
	No preference	19	25.7	4	21.1	23	24.7	
	Read units of measure	24	32.4	9	47.4	33	35.5	
If you had to choose one of these ways to have items read to you on a test, which way would you choose?	Do not read units of measure	37	50.0	10	52.6	47	50.5	.28$^{\text{n.s.}}$
	Read units of measure	36	48.6	9	47.4	45	48.4	
	Missing	1	1.4	0	.0	1	1.1	

Note. $^{\text{n.s.}}$ p>.05

Figure 4. Percentage of Students in each Option of Scripting Rule in Tables/diagrams by Disability Status

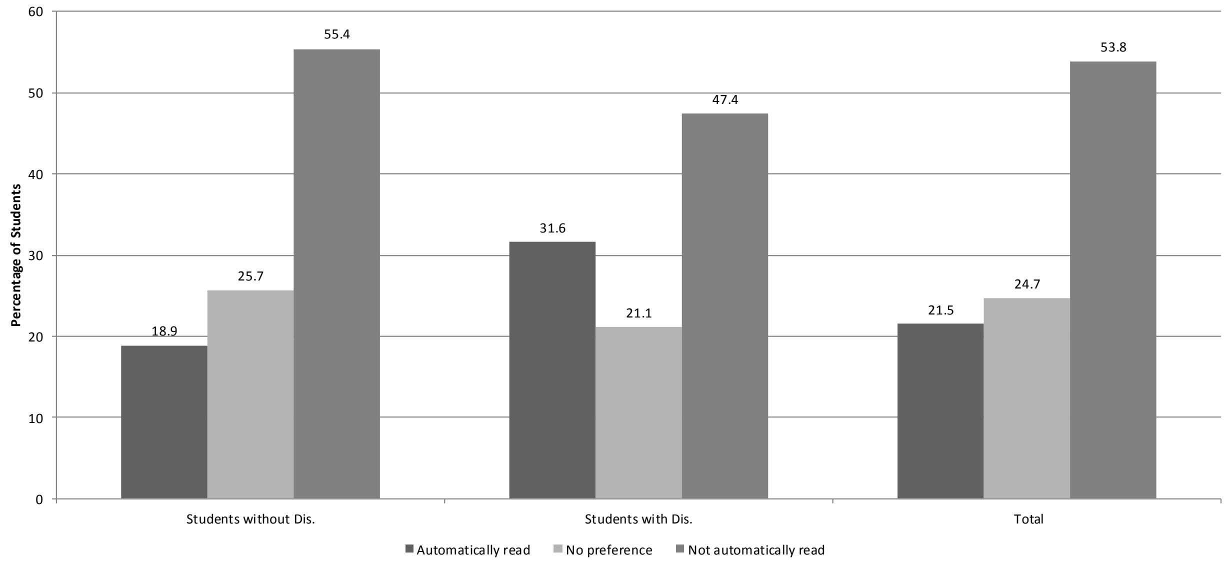


Figure 5. Percentage of Students in each Option of Scripting Rule in Chemical Equations by Disability Status

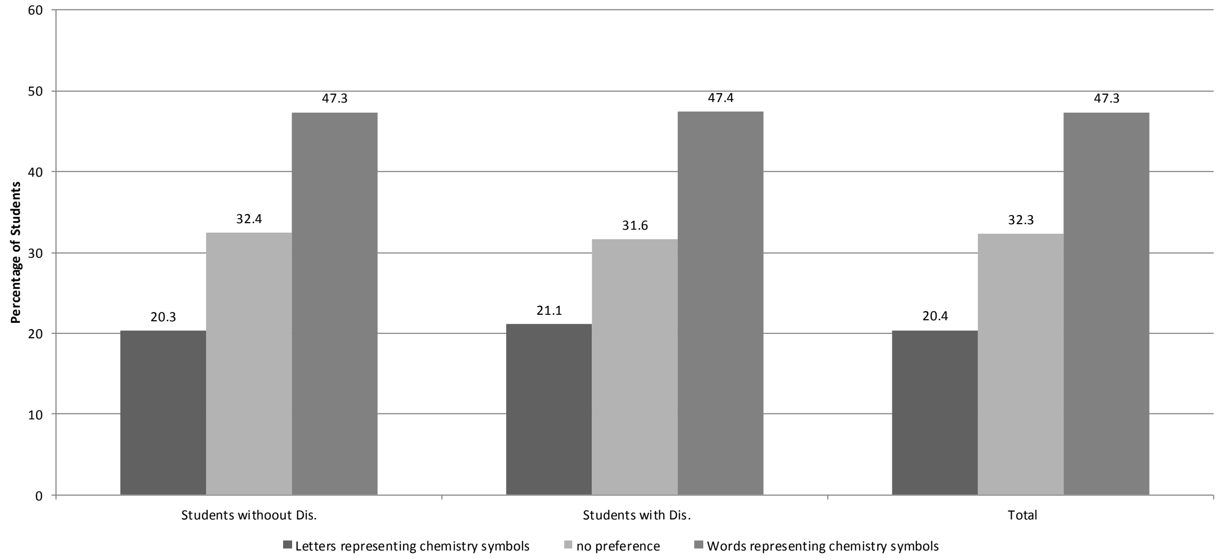
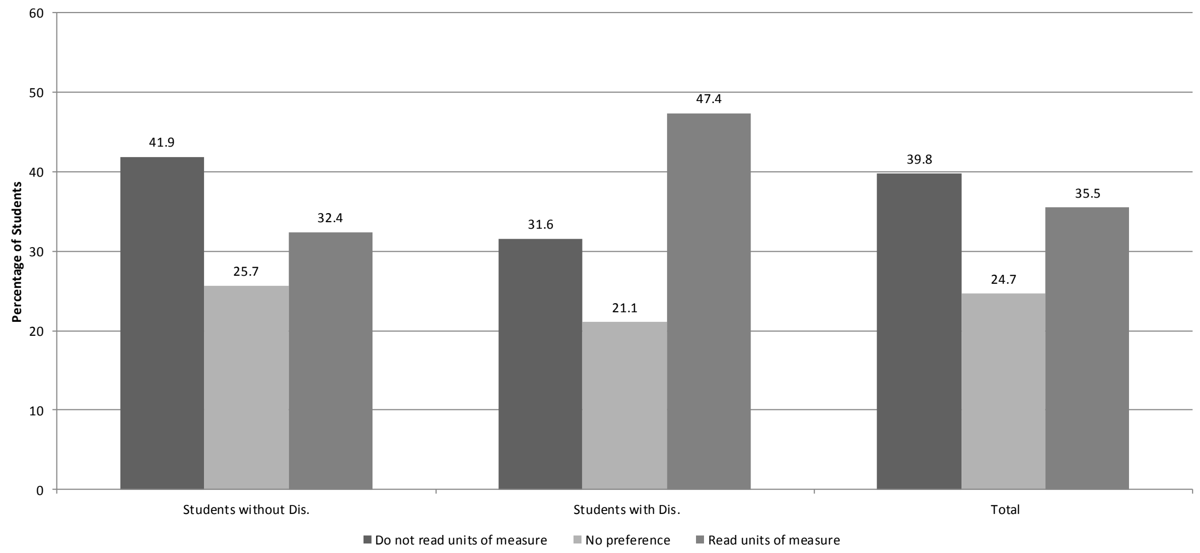


Figure 6. Percentage of Students in Each Option of Scripting Rule in the Units of Measure in a Table by Disability Status



Summary of Results

Quantitative Results

Although the number of items and sample size for students with disabilities was too small to yield results that were statistically significant, descriptive statistics provided some information about student performances and preference. Overall, there was an achievement gap between students without disabilities and students with disabilities. Further, there was not a statistically significant difference in performance by scripting rule (i.e., scripting rules did not impact student performance). However, on a descriptive level there were differences worth noting. For example, students with disabilities' mean scores were .13 higher on items in which tables were read automatically than when they were not. Mean scores for students with disabilities were .21 higher for items in which chemical equations were read as letters than when items were read as words, and mean scores were .05 higher when units of measures were read in tables than tables that did not have units of measure read. These findings were the complete opposite pattern of that for students without disabilities, whose mean scores were nearly even (<.03 difference) for tables and chemical equations, and .08 higher when units of measures were not read in tables.

In terms of preferences, a greater percentage of students (both with and without disabilities) preferred that information in tables *not* be automatically read. Both types of students also preferred chemistry equations to be read as words, not letters. The only category of divergence was in tables, where students without disabilities preferred units of measure *not* to be read, while students with disabilities preferred *automatic* reading of units of measure. In two instances (automatic reading of information in diagrams/tables and reading chemical equations as letters/words), students with disabilities' preferences did not align with their performance. In these cases, performance for preferred scripting rules was lower than that of non-preferred scripting rules (on aggregate).

Qualitative Results

Cognitive lab interviews with 16 students were intended to help better understand student views on the presentation options studied as well as general perspectives on their testing experiences. During interviews, students' comments ranged from brief direct answers to specific prompts as well as broader opinions regarding other possible online supports not already provided for them. The open-ended nature of the cognitive interview protocol allowed for students to provide feedback on assessment in general.

The opening interview prompt inquired about student views of the tutorial for using the online assessment accommodation tools. Many students expressed generally positive impressions of the tutorial, in that the steps were clear and comprehensible and the verbal captioning as well as the demonstration of answering test items was helpful for students to understand. Alternately, a smaller number of students indicated that the pacing of the instructional statements and guidance was too slow, repetitious, challenging to one's patience, and difficult to listen to all of the explanations of test-taking steps. Inherent in these comments seemed to be the sense that the test-taking steps were logical and commonsensical enough not to require a procedural explanation. Some comments related to the tutorial's limitations were relevant to students' previous knowledge of using various computer-based supports in the instructional and assessment environment. A convergence of comments yielded the suggestion that it ought to be possible to skip over the verbalization of the instructions. A divergent and unique idea was that the instructions could be embedded into the test itself, so that students actually would be presented with directions as the test items were presented. The point of this idea is that students would be briefly familiarized with the procedures in the context of the test, rather than as a separate introductory segment before the test.

The first set of support alternatives offered the options of automatic reading of all information in the tables or not reading the information in the tables. (See Table 6 for a summary of student preferences about accommodation options.) Across the participant group, there was no convergence of opinion about which was the preferable option: 8 of the 16 preferred automatic reading of all information in the tables, whereas 6 of the 16 preferred no automatic reading, and 2 students had no preference. The chief reason in favor of automatic reading was because this accommodation facilitated comprehension of the item. The chief reason in favor of not automatically reading table information was that the repetition of elements in each cell was perceived as distracting or confusing. To this point, some indicated that the reading of the placeholder zeroes seemed monotonous and unnecessary; for example,

> *. . . I wish they would read one-point-zero-zero [1.00] as just one [1]. For wax they could just say point-nine-five [.95] instead of zero-point-nine-five [0.95]. . . .*

An additional point appears related to the underlying sense of test-taker autonomy: students expressing the preference for the table not automatically being read aloud indicated that they desired to have a choice as to which parts, and how much, of the table contents would be read to them. Their point was that they did not want to deal with things being read which were not necessary for them individually. Their awareness of this tension between their needs and the universal need for oral presentation of the table contents was evidenced by the qualifying statements, "I would not need this, but others might" and "If this were just about me, I would not want this [item version], but I guess some people could need it."

Table 6. Students' Preferences about Accommodation Options and Related Reasons

Scripting Rule	Preferences (of n=16)	Reasons
Audio representation of table contents	8 preferred automatic option	Facilitated data comprehension, in order to understand what item was asking
	6 preferred non-automatic option	Simplified information presented, so as to prevent distraction or confusion
	2 reported no preference	
Audio representation of chemical equation	2 preferred elements read as symbols/letters	Provided for simplicity and clarity of each data point
	12 preferred elements read as words	Facilitated comprehension of meaning of chemical symbols and their relation to test item
	2 reported no preference	
Audio representation of units of measure in cells of tables	6 preferred units read aloud	Facilitated comprehension of data and their relation to test item
	2 preferred units not read aloud	Provided for clarity and simplicity of each data point
	8 reported no preference	

The second set of scripting options provided two ways that chemical equations were presented in auditory form: as series of letters that symbolize the elements that form the equation, or as the full name of each element in the equation. There was a strong degree of agreement regarding the participant group's preference: three-fourths of them indicated that they preferred to hear the elemental symbols and operations read as words. The chief reason reported was that the interpretation of, for example, "Na" as "sodium" and "+" as "reacts with" allowed for students to understand the meaning of the chemical equation and to relate that information to the question being asked in the test item.

Put another way, students thought they neither needed to decode the symbols nor hold the code information in active memory in order to develop an answer to the test item. Alternately, for those who preferred the reading of the equations as letters and numbers, they indicated that when the equations were interpreted using words, it became confusing and unclear to listen to the reading of the test item. Additionally, students observed and commented on another difference: one version of the item presented text about what was expected of test-takers before the table was read, priming students to understand the nature of the test item before related data were reported. Most students who noticed this difference between item versions also indicated that they preferred this version with preliminary text. Other comments repeated the desire for items to be read more quickly (speed) and with fewer pauses, as pauses contributed to distractedness for some test-takers. One student suggested that when test items relate to elements, there could be a Periodic Table available as a clickable "pop-up" to which students could consult.

The final set of presentation options for the reading of the contents of a table offered that the units of measurement would be read aloud or that only the numbers in each cell would be read. There was no strong convergence across the participant group about the preference regarding these options. In this case, half of the group expressed no preference between these two ways of reading the chemical equation. In fact, many of these students actually had not noticed any discernible difference and, when informed of the distinction between the versions, indicated that this difference was unimportant. Of the remaining students who had a preference, most of them preferred that the entire content of the table cells would be read, with the numbers and the units of measure. The chief reason offered was that hearing the measurement units aided in comprehending data.

Alternately, for those who preferred the omission of the units of measure, the chief reason was that this approach would simplify the table and would also save time. Some participants indicated that the reading of numbers in a table is not necessary for them since they do not have difficulties reading numbers for themselves. A unique comment was that there is a threshold of complexity below which the reading of tables is not necessary; that is, when the table has few rows and columns, the contents ought to all be read aloud, and when the table has many columns and rows, the reading of all information in each cell may not be helpful and may be overwhelming.

Students also made other comments besides their preferences between the three sets of options presented. Multiple comments clustered around reading mechanics: that the reading speed was too slow in general, that the intonation and pacing did not modulate but the speech seemed monotone, and that the pauses seemed unnatural in comparison to human speech. Additionally, some comments noted that the highlighting of words as they were read was helpful in tracking the words visually while hearing them, which improved focus; however, one student indicated that the highlighting was distracting to his attention to the text. Overall, students who were observed to have proficiency using computers for taking the assessment used the word "annoyed" and "annoying" frequently when sitting through segments of the tutorial or through segments of items that took a perceived long time for the scripting to read through them.

Participants in the interviews seemed willing to offer additional feedback and suggestions about things that they would hope to be changed about the assessment. It seems likely that these comments were elicited due to the nature of the cognitive interview process, as the intent is to ask open-ended questions and prompts to allow for stream-of-consciousness communication between participant and researcher. Also, these comments seemed associated with their awareness that the final version of the assessment is not yet finalized. Further, the relative amount of previous computerized test-taking experience of the student participants seemed to encompass a variable that influenced how they perceived the scripting as it was demonstrated in this study.

Proficient computer users, who seemed to be a large portion of the participant group, offered several comments about functions that they would have liked to see in the available tools and capabilities of the online assessment in this study. For instance, suggestions included a desire for such features as calming background music, variable speed controls for the recorded speech, and multiple pitch settings as choices such as lower male tones or higher female tones. Many of these features are currently impossible with natural human speech functions (they are currently only available with word-by-word reading functions).

Other unique comments brought up by one or two participants included the overall approach of how tables or figures were orally scripted, such as aspects of the sequencing and segmenting of the information in the tables. A participant suggested that the headers could be read as a frame for the data, but the actual numbers would go unvoiced. In other words, the table would be structured orally, such as indicating that "the first column is data on …" or "the second row pertains to measures for this thing." Another suggested that it would have been more understandable if the table could have been read by column, left-to-right, rather than by row, top-to-bottom. Another set of suggestions pertained to how tables would be described and orally formatted. For instance, a participant indicated that headers of columns or rows could serve as repeated placeholders, verbally cueing test-takers about the relation of the data to each header. A different opinion was offered where segments of tables could be read together in a sort of structural sentence, such as "This thing has this and this measures, whereas this other thing has that and the other measures." In this case, the student provided a strategy for comprehending the interrelationships among the cells in the tables.

Overall, interview participants stated consistently that they had a positive experience with the computerized assessment, with its oral administration capacity, in comparison to previous testing experiences. They indicated that they were excited at the prospect of using tests with these accessibility features in the future. The students' energy in attending the interview sessions, and their earnestness in offering their ideas and perspectives, may be attributed to their valuing of these types of assessments administered online.

Discussion

This multi-method study was designed to help researchers and policymakers better understand how to "script" an assessment with an audio function. The assessment creators (Nimble Assessment Systems, now Measured Progress Innovation Lab) took care to identify reasonable scripting possibilities, and then assessed students for both performance and preference using both qualitative and quantitative methods.

Quantitative methods, in terms of statistical significance, were inconclusive in this study. There was only one notable statistically significant finding, which identified that students without disabilities performed at a higher rate than students with disabilities on chemical equation items. This finding confirms previous research, but did not answer any new questions about scripting.

The challenge in reaching statistical significance through inferential methods may be due to the total number of items found in each category of scripting rules. In this study, only three items were provided for each scripting rule. Thus, the difference between groups was not sensitive enough to be detected. Second, although students were randomly assigned, there was not a pre-test, which would have helped to understand (and control for) group differences in this study. Finally, the target group in the study (students with disabilities) was only represented by a sample of 19 students. This sample was too small to produce meaningful inferential findings.

The study, however, produced some interesting descriptive and qualitative findings. For example, 47.4% of students with disabilities preferred that information in tables not have automatic audio presentation (31.6% preferred automatic audio presentation). However, students with disabilities scored slightly better when such information was read automatically (students without disabilities' preferences aligned with performance). Likewise, all students preferred that chemical equations be read as words and not letters. However, students with disabilities performed better when equations were read as letters. Finally, students with disabilities preferred that units of measurement be read aloud in tables while students without disabilities preferred no reading (these preferences aligned with performance).

An interesting set of dilemmas emerged from these data. Readers should consider all of these dilemmas with caution, given the small sample sizes present. In two cases there was popular consensus around scripting rules (that information in tables not be read automatically and that chemical equations be read as words). In each of these cases there was a divergence between preference and performance for students with disabilities (this divergence did not exist for students without disabilities). In many states, students with disabilities will be the only population of students who have access to audio representation. However, in states where all students have access to audio representation, this study demonstrated that the two populations (students with and without disabilities) may interact differently with scripting inputs. Further, it is unclear under which condition students with disabilities are best able to show what they know. Qualitative and quantitative perception data leads us in one direction, while performance data in another.

It is clear that more research is needed. Specifically, a larger sample of students with disabilities will be needed to understand the impacts of scripting rules on this population. For students without disabilities, there appears to be a convergence between their preferences for scripting choices (information in tables not automatically read, chemical equations written as words, and units of measure not read). For students without disabilities, performance and preference align, yet students with disabilities had scripting preferences that did not lead to better assessment results. Even from this small study, it may be possible to pilot scripting approaches and track results on a larger scale.

Because the data on students with disabilities was so erratic in this study, pursuing a follow-up study with a larger and more representative sample of this population may be helpful. Given apparent U.S. Department of Education policy shifts eliminating assessments designed specifically for students with disabilities (i.e., U.S. Secretary of Education Duncan's March 15, 2011, press release, stating that the Department is moving away from the use of Alternate Assessments based on Modified Achievement Standards), it is more important than ever to ensure that the unique information comprehension needs of students with disabilities are reflected in assessments. A broader study may confirm the findings of this study, which would mean difficult questions would need to be answered on scripting. On the other hand, a larger sample may demonstrate patterns more similar to students without disabilities. In either case, scripting decisions for students with disabilities appear to be currently in limbo and in need of further study.

References

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). White paper: *Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS)*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Christensen, L. L., Braam, M., Scullin, S., & Thurlow, M. L. (2011). *2009 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 83). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/v3n7.shtml

Ericsson, K. A., & Simon, H. A. (1994). *Protocol analysis: Verbal reports as data (Revised edition)*. Cambridge, MA: MIT Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement (3rd edition)*. Washington, DC: American Council on Education.

Messick, S. (1996). *Validity and washback in language testing*. Princeton, N. J.: Educational Testing Service.

National Assessment of Educational Progress. (2009). *Administer: A manual for assessment administrators*. Washington, DC: National Assessment Governing Board.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy* (Synthesis Report 41).Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Quenemoen, R., Altman, J. R., & Cuthbert, M. (2008). *Trends in the public reporting of state assessment data (2001-02 through 2004-05)* (Technical Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Bremer, C., & Albus, D. (2011). *2008-09 publicly reported assessment results for students with disabilities and ELLs with disabilities* (Technical Report 59). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix A

Examples of Content Elements and Scripting Rules for Each

Tables/Diagrams

Element X reacts with potassium (K) to produce the compound K_2X. The table below shows the number of valence electrons in four elements.

Valence Electrons in Four Elements

Element	Number of Valence Electrons
Hydrogen (H)	1
Nitrogen (N)	5
Oxygen (O)	6
Fluorine (F)	7

Which element listed in the table is **most likely** element X?

A. hydrogen
B. nitrogen
C. oxygen
D. fluorine

Scripting Rule 1:

1. Read first two sentences.
2. Read Table Title and all table elements.
3. Read question.
4. Read answer choices.

Scripting Rule 2:

1. Read first two sentences.
2. Read table title.
3. Read question.
4. Read answer choices.

Scripting Rule 3:

1. Read first two sentences.
2. Move question above table and read after first two sentences.
3. Read table title and all table elements.
4. Read answer choices.

Chemical Equations

$$2C_2H_6 + 7O_2 \longrightarrow 4CO_2 + 6H_2O$$

In the combustion of ethane, how many moles of CO_2 can be produced from 1.00 mole of C_2H_6?

A 0.500 mole

B 1.00 mole

C 2.00 moles

D 4.00 moles

Scripting Rule 1: Read equation as "two ethane molecules combine with seven oxygen molecules react to form four carbon dioxide molecules and six water molecules"

Scripting Rule 2: Read equation as "Two C two H six plus seven O two react to form four C-O two plus six H-two-O."

Units of Measure in a Table

The table below shows data from a heating experiment.

Metal	Heat Added (J)	Mass of Metal (g)	Change in Temperature (ºC)
copper	3000	100	77
iron	3000	100	64
lead	3000	100	231
silver	3000	100	130

Which of the following conclusions is supported by the data in the table?

A. A given mass of silver requires less heat to change its temperature 1ºC than an equal mass of iron.

B. A given mass of silver requires less heat to change its temperature 1ºC than an equal mass of lead.

C. A give mass of copper requires less heat to change its temperature 1ºC than an equal mass of lead.

Scripting Rule 1: Read all table elements as they are presented, with units of measure. For example, row one would be read "copper, three thousand Joules, one hundred grams, seventy seven degrees Celsius."

Scripting Rule 2: Read all table elements as they are presented, with units of measure. For example, row one would be read "copper, three thousand, one hundred, seventy seven."

COLLEGE OF EDUCATION
+ HUMAN DEVELOPMENT

UNIVERSITY OF MINNESOTA

NCEO is an affiliated center of the Institute on Community Integration