OXFORD

Journal of Deaf Studies and Deaf Education, 2016, 383–393

doi:10.1093/deafed/enw051
Advance Access publication August 19, 2016
Empirical Manuscript

EMPIRICAL MANUSCRIPT

Development of American Sign Language Guidelines for K-12 Academic Assessments

Jennifer A. Higgins, Lisa Famularo[1], Stephanie W. Cawthon[2], Christopher A. Kurz[3], Jeanne E. Reis[4], and Lori M. Moers[5]

[1]Research Matters, LLC, [2]The University of Texas at Austin, [3]Rochester Institute of Technology, [4]The Center for Research and Training at The Learning Center for the Deaf, and [5]Maryland School for the Deaf

Correspondence should be sent to Jennifer Higgins, 2135 North Pantops Drive, Charlottesville VA 22911 (e-mail: jcowan80@hotmail.com).

Abstract

The U.S. federal Every Student Succeeds Act (ESSA) was enacted with goals of closing achievement gaps and providing all students with access to equitable and high-quality instruction. One requirement of ESSA is annual statewide testing of students in grades 3–8 and once in high school. Some students, including many deaf or hard-of-hearing (D/HH) students, are eligible to use test supports, in the form of accommodations and accessibility tools, during state testing. Although technology allows accommodations and accessibility tools to be embedded within a digital assessment system, the success of this approach depends on the ability of test developers to appropriately represent content in accommodated forms. The Guidelines for Accessible Assessment Project (GAAP) sought to develop evidence- and consensus-based guidelines for representing test content in American Sign Language. In this article, we present an overview of GAAP, review of the literature, rationale, qualitative and quantitative research findings, and lessons learned.

Including all students in state assessment programs, particularly students with disabilities and English learners, is critical for equality and "has generated considerable and sometimes frenzied activity during the past two decades" (Thurlow & Kopriva, 2015, p. 1). In the United States, the Every Student Succeeds (2015) and Individuals with Disabilities Education Improvement Acts require state testing programs to provide appropriate accommodations or "changes in assessment materials or procedures" (Thurlow & Bolt, 2001, p. 3) for students during testing. A wide variety of accommodation options are available during state testing, and it is the responsibility of teachers and other school personnel, with student and parent input, to decide which accommodations are best fit for individual students' needs on a given assessment in order to allow students to best demonstrate their academic proficiency (Christensen, Braam, Scullin, & Thurlow, 2011). Without access to appropriate accommodations, many students are placed at a disadvantage in demonstrating their proficiency (Madaus, Russell, & Higgins, 2009; Thurlow, Elliott, & Ysseldyke, 1998). Although deaf or hard of hearing (D/HH) students may use an American Sign Language (ASL) accommodation during testing, little information exists on how best to translate English test items to ASL and few research studies have been conducted to examine the impact of the ASL accommodation on students' performance.

The Guidelines for Accessible Assessment Project (GAAP) sought to address this need by engaging with partners from 18 state departments of education and a team of experts in academic ASL and inclusive assessment. The GAAP team iteratively developed evidence- and consensus-based guidelines for creating test items in ASL and corresponding exemplar ASL representations of test items aligned with college and career ready learning standards. This article presents key project findings and includes: (a) a review of state practices on the ASL accommodation; (b) findings from cognitive labs with students to explore the impact of different ways of representing test content in ASL form; (c) findings from a randomized controlled trial (RCT) to evaluate the effects of ASL supports and different representations on students' performance; and (d) lessons learned from project meetings and research activities.

Received May 17, 2016; revisions received July 8, 2016; accepted July 11, 2016

ASL Translation as an Accommodation

There are no known research studies on guidelines for signed representation of academic test content and few research studies published on the impact of sign accommodations on K-12 students' performance. Most of the existing research related to ASL as an accommodation for instruction or assessment has been conducted in a postsecondary education environment with a high enrollment of deaf students: Rochester Institute of Technology in Rochester, New York. Marschark, Sapere, Convertino, Seewagen, and Maltzen (2004) published early work that investigated the effects of ASL interpretation with transliteration, or "English-like signing" instructional accommodations on students' performance, as well as potential interactions with assessment delivery format, either written or signed. This study found that there were no effects of either the different types of signed presentation of material on outcome measures or the interactions with assessment delivery format. In a follow-up study, Marschark, Sapere, Convertino, and Seewagen (2005) measured the effects of several additional interpreter-related variables on students' performance to investigate potential nuances that may affect student learning in an accommodated instructional environment (assessments were not interpreted in this study). More specifically, this study included measures of students' preferences for interpreting versus transliteration, familiarity between the student and the interpreter, and interpreters' experience; results from this analysis indicated that these variables were not shown to have an effect on students' performance on study posttests. Finally, in a third study, Marschark, Pelz, Convertino, Sapere, Arndt, and Seewagen (2005) examined the effects of live presentation of ASL in instruction versus when it was presented in a video. Measures included information both about students' performance and about their allocation of visual attention across multiple sources of information. This study found that students' performance on study outcomes (again, not accommodated) does not differ when ASL interpretation of academic content is provided via live interpretation versus ASL video-taped interpretation.

At the K-12 level, research on the use of an ASL accommodation for assessments provided evidence of no statistically significant difference when students were provided sign support versus no support (Cawthon, Winton, Garberoglio, & Gobble, 2011). One factor that may have contributed to the nonsignificant result is familiarity of the ASL translation to the student. In this study, the form of ASL used in the test translation was described as being "different from conversational and instructional ASL," instead focusing on a conceptually accurate ASL representation that closely followed the meaning and structure of the test item (Cawthon et al., 2011). Furthermore, because students had the option of reading the English text in addition to or instead of using the ASL representation (i.e., English text was not removed all together for student responses), the consistency in which the students used the ASL accommodation and corresponding impact was unclear.

There is limited research available that provides information about students' preferences for digital delivery of ASL videos, and articles that do mention this topic provide little information about reasons for students' preferences. One study collected feedback from students and test administrators on test format in an effort to compare D/HH students' performance on paper and pencil based mathematics test items with and without ASL support (Maihoff et al., 2000). To ensure uniformity in ASL support, it was delivered via a DVD of a human signing the test items. The article reports that all students stated that the ASL items on DVD were "easier to understand" than the paper and pencil based test items. Despite this, some students preferred the written version because it took less time to complete. The DVD and paper format added to the overall test time because all students were required to view a common DVD, rather than each student having control over pacing via their own DVD. The test administrators also reported that the ASL presentation of test items on the DVD was clearer and an improvement over ASL presentation by individual teachers during an assessment (Maihoff et al., 2000). The authors did not report any details about why students believe the DVD of signed test content was "easier to understand" nor why the presentation of test items on DVD was an improvement. In a more recent study, researchers compared students' performance and collected student feedback on ASL items delivered in two formats via a computer-based testing system: videos of a human signing and videos of an avatar signing (Russell, Kavanaugh, Masters, Higgins, & Hoffmann, 2009). The study found that although students' performance and test taking time did not significantly differ, two thirds of students preferred the human signer over the avatar. Although the students' overall reactions to the computer-administered signed representation of the test were positive, specific reasons for students' preferences were not collected as part of the study.

Current Policies and Practices

In order to document and learn from existing state ASL assessment practices, researchers searched the websites of 50 state departments of education and collected state accommodation manuals, test administrator manuals, and documents related to ASL delivery of assessment content. After collecting relevant documents, researchers analyzed all text related to the development and delivery of content in ASL to identify the most common practices and policies/practices that were not consistent across states. Analysis of state documents related to sign administration of assessment yielded four key themes:

1. Many states had rules and regulations for the qualification of individuals who sign test content for students (i.e., interpreters) and stressed the importance of the interpreters' familiarity with test content and terminology. Some states suggested that the student's teacher acts as the interpreter for assessment content, some states allow it only if a second interpreter is present to monitor interpretation, and some states forbad it. Although this finding is not relevant to the digital delivery of sign representations of test content because the signed representations are developed a priori by test developers, it provided evidence that sign guidelines are in their infancy and there is no common agreement among states.
2. States warned against cueing/cluing, elaboration, and clarification and provided direction for using non-manual markers (e.g., facial expressions, body language, and objects), fingerspelling (i.e., the process of presenting each letter of an English word or term individually), writing and pointing to content on the board, and interpretation of graphics. Many states explicitly stated that math symbols and terminology must be fingerspelled in order to ensure that additional construct-relevant information is not being provided to students receiving the signed version of the item. For example, the sign for parallel lines is two index fingers parallel to one another and thus, the concern is that the sign shows students what it means for two lines to be parallel. On the other hand, Texas policy stated that if a sign for a word or phrase exists, the test administrator should use the sign when the word or phrase is used in the English text version of the test. Texas policy stated that fingerspelling is not an acceptable substitution because it increases the difficulty of the item by requiring the student to recognize a

term by its spelling and given that hearing students would not be required to recognize a term by its spelling in an oral administration, it should not be required of a student who is deaf (Texas Education Agency, 2012).

3. Three states had guidelines for the use of nonstandard signs or "locally developed signs." These signs are known by the student and sign administrator, but are likely not standard across schools or states. Texas sign guidelines provided "fission" as an example of a word that may have a locally developed sign, which could incorporate the concept of "splitting apart" (Texas Education Agency, 2012).
4. Many states had guidelines for the testing environment (lighting, size of student group, etc.), logistics of administration (scheduling of interpreter, viewing of test materials in advance of testing, etc.), and test security procedures and guidelines.

The literature and state document reviews were critical steps in the identification of specific issues to focus on in the cognitive labs and randomized control trial and further emphasized the need for research and development of sign guidelines for state test content.

Challenges to Validity

Administering standardized tests in English print may create a barrier for students who receive instruction in a language other than English, whose primary language is not English, or who cannot access print (e.g., students who are blind). For this reason, state test content is often translated into multiple forms, most often when the goal is to measure students' proficiency in content areas other than English (e.g., social studies, mathematics, and science). It is critical that different linguistic representations of test content (e.g., braille, Spanish, and ASL) remove the construct-irrelevant barrier of English and allow students to show what they know and can do, ultimately providing a more reliable and valid measure of student proficiency in the assessed content area. D/HH students who receive instruction in ASL, or whose primary language is ASL, are often offered an ASL accommodation on a given test in order to provide greater access to test content.

The most common approach to providing ASL or signed support to students during testing involves a teacher or interpreter translating the test content for a student during testing. This approach is problematic because it introduces uncontrolled variability into what is meant to be a standardized test. Exacerbating the problem further, school personnel are often not given ample time to review test content prior to test administration due to concerns about keeping test content secure, and thus, are required to translate the content at the time of testing. Due to variations in language use, there may be a variety of ways that content can be presented when live ASL interpretation, sometimes referred to as "translation on the fly," is employed. This introduces variability in the delivery of test content to students. One research study found that local translations of test content are inconsistent across administrations and vary in quality, negatively impacting the validity of the inferences made from the assessment results (Qi & Mitchell, 2012).

Standardization in large-scale assessment is a critical factor in ensuring that tests are reliable and scores lead to valid inferences about students' skills and knowledge (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). To ensure that scores on tests translated to ASL are comparable with scores on English print tests, it is imperative that ASL test items have equivalent item characteristics (e.g., item length and difficulty) and measure the same construct. This means that different factors should be considered and different expertise is necessary when translating a mathematics test versus an English Language Arts test. In a study that investigated the impact of translating mathematics assessment items into ASL, researchers analyzed recordings of teachers signing test items as they would to their students. They found that variations in the translations and features of sign language affected item difficulty (Ansell & Pagliaro, 2001).

Although inconsistencies exist whenever an educator or other person (e.g., sign interpreter, human reader, or scribe) provides an item translation or another type of accommodation, ASL translation is particularly problematic due to the lack of well-trained, highly skilled, linguistically and culturally fluent educational interpreters. Research on interpreter quality suggests that more than half of all D/HH students are assigned interpreters with inadequate skills to provide sufficient access to academic content in the classroom (Schick, Williams, & Kupermintz, 2006). This means that students assigned to a less qualified interpreter during assessment are likely to have a disadvantage compared with students assigned to a highly qualified interpreter. In sum, these factors contribute to a substantially different experience for D/HH students taking standardized tests via ASL "translation on the fly," an experience that deviates significantly from standardized test design principles.

ASL Accommodation Innovations

Recent advancements in computer-based assessment enable the ASL test accommodation to be delivered digitally, thus eliminating many of the issues associated with live translation. Specifically, technological advances make it possible to embed videos in a test delivery system, allowing the test item in English text and ASL video to be displayed simultaneously. With this online delivery system, students can read the test item in English, view the ASL item on video, with full control over both. Delivery and access to both are equal: the English item can be re-read, and the ASL item replayed, in entirety or in sections. This approach is often referred to as "embedded ASL video support." The design allows each student to individually decide how to access the test content in a way that best fits his or her needs, giving D/HH students similar delivery characteristics and options as their peers.

The change in delivery mode from human to computer requires a change in responsibility for developing and delivering ASL versions of test items. Responsibility shifts away from schools and districts who have traditionally required school personnel or contracted interpreters to sign test content to students, to test developers who create the ASL test items a priori. Specific rules and criteria for developing ASL test content are needed in order to maintain the reliability and validity of inferences made based on student scores. Furthermore, research is needed to ensure the guidelines result in ASL test items that remove construct-irrelevant barriers for students and lead to valid inferences about their knowledge, skills, and abilities without violating measurement of the construct of interest. The GAAP, an initiative funded by the U.S. Department of Education, sought to address these needs.

Current Project

The GAAP ASL team worked together to identify areas where research was needed in order to decide on an appropriate

guideline for ASL delivery of test content. Five specific issues were identified and subsequently studied via cognitive labs and a RCT: (a) presentation of mathematical notation and other images, (b) fingerspelling key terms, (c) item structure (diamond), (d) plurality, and (e) use of space.

Presentation of mathematical notion and images

Presentation of mathematical notion (e.g., expressions and equations) and images was selected as an issue to investigate because the GAAP team questioned the rationale for the guidelines in this area provided in state documents. Several state documents advise interpreters to point to content such as graphs and equations as a way to refer to this information rather than presenting the content in ASL. The GAAP team questioned whether students need ASL access to the English words only or whether they also need signed access to mathematical notation and graphics. Two concerns about signing math notation and images were raised by the ASL experts on the GAAP team. The group was concerned that presenting these types of content in ASL might be cognitively cumbersome. For example, presenting a complex and lengthy equation (e.g., an equation that includes several parenthetical expressions and multiple mathematical operations) or presenting information on a data-heavy graph (e.g., a line graph that shows trends over 10 years for five countries) requires many ASL signs that the student might not need in order to understand the test item. The second concern was that the student could be confused if the ASL item did not present the mathematical notation or graphic in a manner consistent with the student's instruction. Due to the fact that instructional practices and fluency in ASL vary widely across schools and classrooms and the presentation of this information could be complex, the GAAP team deemed the presentation of mathematical notion and graphics worthy of study.

Fingerspelling of key terms

Test items often include specific key terms that are integral to the construct being measured. Many states' sign guidelines explicitly state that some terms need to be represented to D/HH students via fingerspelling because there is a concern that signing these terms may cue students to the correct answer and/or provide extra information. However, as noted earlier, Texas policy recommends limiting the use of fingerspelling due to the potential of increasing item difficulty. Fingerspelled terms are cognitively complex, as they represent English words letter by letter. The GAAP team evaluated how presenting terms in ASL, and via the English-based system of fingerspelling, could provide similar and different information based on the context and what is being measured. Fingerspelling key terms in test items was studied in order to understand students' preferences and the impact of fingerspelling on performance.

Item structure (diamond)

The GAAP team selected three issues related to the linguistic conventions of ASL. One area of exploration was whether the structure of the test item should follow ASL discourse conventions rather than English-based conventions. The English-based linguistic structure of some test items is to start with general information, often in the form of a short narrative passage or a table containing data. This is followed by a question. GAAP researchers hypothesized that items composed to align with the ASL discourse strategy referred to as "diamond structure" might focus D/HH students' attention and engagement with the content in a linguistically appropriate manner. Items set up using the diamond structure introduce the question or goal first and restate it at the end, with item information in between. The rationale for this hypothesis was that ASL pragmatics sometimes requires that the main reference point, topic, or goal of the discourse be established at the start and end of longer ASL texts, and associated details discussed between those two points. This structure also clarifies the goal of the test item first, perhaps creating a context for the remaining components of the item.

Plurality

A second linguistic issue GAAP researchers studied was plurality. Many test items, particularly in lower grade level mathematics, contain verbs that represent repeated action or plural nouns. GAAP researchers were interested in understanding more about the impact of representing repeated actions and plural nouns using different linguistic mechanisms that are seen in ASL. The team composed items in which plurality was expressed via ASL number terms, reduplication for specific number, and reduplication for general plurality with no specific number, for example, mass nouns. In the ASL number term structure, an action or item was presented in singular form and a number sign added on to indicate the number of objects or repetitions. In the reduplication structure, the root term or a related stem was signed with a repeated or sweeping motion to represent the same object as plural, or the same action as occurring more than once. Plurality was studied in order to understand students' preferences and the impact of these two processes on students' performance.

Use of space

The third linguistic issue studied by GAAP researchers was use of space. When composing ASL items, it is important to consider the relationships between the spatial locations of various elements, in order to present them accurately and consistently. Items that are not carefully crafted in this regard are more likely to misrepresent the content. When referencing an item element such as a graph, the signer has the option of referring to the graph in general space (in front of the signer's body) or specific space (on the signer's hand) as it would appear on a monitor in the signer's point of view. Although either option is acceptable in ASL, GAAP researchers were interested in studying whether one version was preferred by students or whether students performed better with one of the two versions.

Each of the five issues described was studied in the cognitive labs and in the RCT. Some issues were studied across all grade levels, whereas some were studied in only one grade-level band. GAAP researchers chose which issues to study at each of the grade-level bands by reviewing item content and characteristics.

Methods

GAAP used a mixed methods research design to develop evidence- and consensus-based guidelines for creating ASL representations of test items and corresponding exemplar test items. Research methods included cognitive labs with students to explore the impact of different ways of representing test content in ASL form and a RCT to evaluate the effects of computer-embedded ASL accommodation and different ASL representations on students' performance.

Cognitive labs are face-to-face interactions during which a researcher observes and evaluates a student's cognitive processes. Cognitive labs have become a widely used method of gathering evidence related to the validity of inferences made by assessments, specifically evidence about whether assessment

items are measuring the intended constructs (Dolan, Goodman, Strain-Seymour, Adams, & Sethuraman, 2011; Ericsson & Simon, 1999; Gorin, 2006; Willis, 1999). For this research study, a cognitive lab method that included structured prompting to guide students in articulating their thoughts was employed. All cognitive labs were conducted in schools for the deaf. During each cognitive lab session, a student worked with two hearing researchers using an ASL–English interpreter who had experience working with students in an education setting and who was recommended by the school contact. The researchers first explained to students the purpose of the study and explained that they would be asked to complete a series of test items using a computer-based testing system with ASL provided via a video of a human signing the test content, followed by a short interview. The student was then presented with four or five pairs of mathematics and English Language Arts items. Each item pair illustrated two different ways that test content could be presented in ASL. To isolate the effect of the presentation under investigation, all other characteristics of the two items were parallel, meaning they differed only in surface characteristics. As an example, the item in Figure 1 asked the student to select a sentence that best describes the shapes that are presented. The ASL video for this item showed the signer presenting the terms "right angle," "line of symmetry," and "length" using the ASL sign only.

The second item in the set was an isomorph or parallel item, meaning the item differed only in surface characteristics such as the shapes that are presented and the name of the person in the item. In the second item of this pair, the signer presented the terms "right angle," "line of symmetry," and "length" by fingerspelling the English term letter by letter followed by the ASL sign. These were both conceptually key terms and terms which may not have standard ASL equivalents across schools. Although the task itself was the same in the two items and the mathematical concept involved was very similar, the presentation of key terms, selected and determined by the GAAP team, differed (Figure 2).

For each item, students were encouraged to view the video and respond to the item before moving to the next item in the pair. After completing the pair, students were then asked questions about the two items, including the clarity of the information presented in the ASL video and whether there was anything confusing; students were also asked to report a preference for either the item with the ASL sign only or the item with both a fingerspelled English term and ASL sign in sequence. Researchers took notes on whether the student struggled with test content and whether different item features were familiar to the student. At the end of the cognitive lab session, researchers (via the interpreter) asked for any feedback on the ASL videos used during the cognitive lab session and collected student background information, such as at what age he/she learned ASL and whether the student had other deaf family members.

A separate phase of the study involved a RCT conducted to investigate the effect of providing computer-embedded ASL support during testing (the intervention) on students' performance on an assessment (the outcome). The researchers also sought to examine the impact of different ASL representations of test items on students' performance. Students completed a three-item orientation to become familiar with the computer-based testing environment and embedded accessibility support and a 19-item mathematics test. Mathematics items were used because ASL translation is more often available on state tests in mathematics than English Language Arts. The mathematics test was administered for research purposes only and consisted of released state and consortia items that were publicly available and aligned with college and career ready learning standards. A variety of item types were employed: multiple-choice, constructed-response, technology enhanced (e.g., drag and drop where students were asked to move objects from one area to another to respond, hot spot where students were asked to click on an area of the items physical space such as a phrase in a sentence or bar on a bar chart to respond).

The RCT was conducted with D/HH students who normally receive ASL support for assessment. A stratified random sample design was employed with teacher rating of D/HH students' mathematics and reading ability forming the strata. Students within each stratum were randomly assigned to one of three test forms. Each test form consisted of the same 19 items, in the same order, but with different versions of ASL support. Participating students received two blocks of items with the intervention (one block with Support Variation 1 and one block with Support Variation 2) and the control condition (block of items with no support). Table 1 shows the different ASL versions applied to test items for the three issues studied at grades 3–5: fingerspelling, use of space, and plurality.

The same form design was used for grades 6–8 and 9–12. At these two grade levels, the issues studied were fingerspelling (Variation 1 key terms were fingerspelled only for Grades 6–8 and fingerspelled and signed for grades 9–12, Variation 2 key terms were signed only for both grade-level bands), presenting equations (Variation 1 equations were signed, Variation 2 equations were not signed), and item structure/diamond (Variation 1 used diamond structure, Variation 2 did not use diamond structure). On

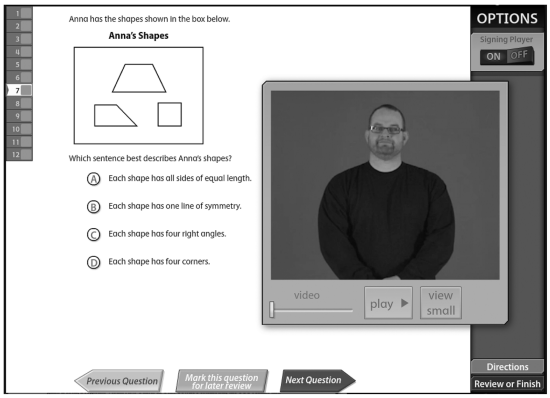

Figure 1. Cognitive lab Item 1 example.

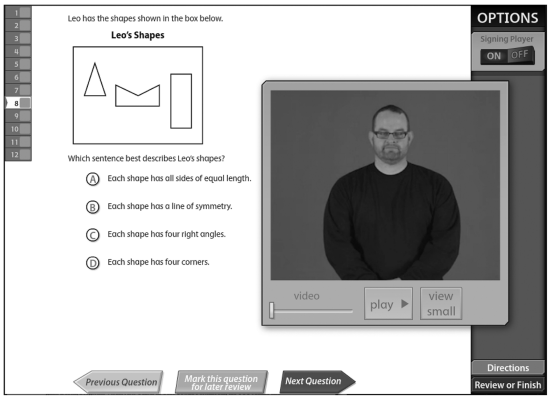

Figure 2. Cognitive lab Item 2 example.

Table 1. Grades 3–5 form design with issues studied and American Sign Language (ASL) versions

	Issue/Item	Form 1	Form 2	Form 3
	1. Introductory Item	ASL	ASL	ASL
Block 1	2. Fingerspelling Item 2	ASL fingerspell only	ASL fingerspell+sign	No ASL
	3. Fingerspelling Item 5	ASL fingerspell only	ASL fingerspell+sign	No ASL
	4. Use of space Item 2	ASL general space	ASL specific space	No ASL
	5. Use of space Item 5	ASL general space	ASL specific space	No ASL
	6. Plurality Item 2	ASL no reduplication	ASL reduplication	No ASL
	7. Plurality Item 5	ASL no reduplication	ASL reduplication	No ASL
Block 2	8. Fingerspelling Item 3	ASL fingerspell+sign	No ASL	ASL fingerspell only
	9. Fingerspelling Item 6	ASL fingerspell+sign	No ASL	ASL fingerspell only
	10. Use of space Item 3	ASL specific space	No ASL	ASL general space
	11. Use of space Item 6	ASL specific space	No ASL	ASL general space
	12. Plurality Item 3	ASL reduplication	No ASL	ASL no reduplication
	13. Plurality Item 6	ASL reduplication	No ASL	ASL no reduplication
Block 3	14. Fingerspelling Item 1	No ASL	ASL fingerspell only	ASL fingerspell+sign
	15. Fingerspelling Item 4	No ASL	ASL fingerspell only	ASL fingerspell+sign
	16. Use of space Item 1	No ASL	ASL general space	ASL specific space
	17. Use of space Item 4	No ASL	ASL general space	ASL specific space
	18. Plurality Item 1	No ASL	ASL no reduplication	ASL reduplication
	19. Plurality Item 4	No ASL	ASL no reduplication	ASL reduplication

all forms, the introductory item was excluded from analysis, and each item of the 18 remaining items was scored either 1 for correct or 0 for incorrect creating a possible form score range of 0 to 18.

Results

Cognitive Labs

The purpose of the cognitive labs was to explore the potential impact of different ways of representing test content in ASL. A total of 46 students from five states participated in the cognitive labs: 16 elementary school, 14 middle school, and 16 high school. Table 2 shows the issues studied at each of the three grade-level bands.

Analysis of the cognitive lab data revealed three themes across the grade levels. First, students preferred items where mathematical notation such as equations and graphics were presented in ASL over items where the nontextual content is not presented in ASL. When asked to express a preference for how nontextual content such as mathematical notation and graphs should be presented, the majority of students across all three grade levels reported preferring ASL presentation of this information. Comments made by students during cognitive labs revealed two primary reasons for the preference. First, students reported that items with mathematical notation signed are more consistent with how this information is presented during instruction. When asked about an item where only text was signed (not the equations), a high school student who is a novice ASL user remarked "Our teacher wouldn't talk about things like that." Furthermore, the omission of signed mathematical notation seemed to confuse some students at the elementary school level. One student, a native signer, described items with equations signed as "more helpful" and indicated he "couldn't follow" the items where equations were not signed. Second, students preferred items where nontextual content such as pictures and graphics were signed because this approach is consistent with ASL as a visual language and how ASL is used both in the classroom and community. One high school student who preferred an item with a picture described in ASL over one where the picture was not described in ASL explained the reason for his preference: "in deaf culture, we do describe photos." Other students made similar comments about preferring nontextual information being signed versus only text being signed with one student explaining that the item was "more like ASL than English."

Table 2. Cognitive lab issues studied

Issue	G 3–5	G 6–8	G 9–12
Equations/Graphic (equation/image signed vs. not signed)	X	X	X
Fingerspelling (key term fingerspelled only vs. signed only vs. fingerspelled and signed)	X	X	X
Item structure (presented in diamond vs. no diamond)		X	X
Plurality (reduplication vs. showing action and referring to number of times)	X		
Use of space (graphic presented on signers hand vs. in front of body)	X		

The second theme that emerged from the cognitive lab data relates to fingerspelling. The cognitive labs studied whether terms should be fingerspelled only, signed only, or fingerspelled and signed in response to two potential issues the team wished to examine: (a) where the ASL term may be unfamiliar to students and (b) where ASL translation of the English term might be disallowed by states and consortia, due to concerns over an ASL term representing the concept in such a way that it may violate the measurement construct. On the second point, GAAP findings are presented in the discussion section of this article. On the first, cognitive lab results provided evidence that students prefer that terms either be signed in ASL, or signed and fingerspelled, as opposed to just fingerspelled. Students who knew the ASL sign for a particular term preferred the item where that term was signed to the item that included the sign and fingerspelling. As one high school student explained, "just the signs was fine." Students who may not have known the sign for a particular term preferred the combination of sign and fingerspelling because "when the word matches the sign is most understandable." A middle school student reported using the fingerspelling to confirm the meaning of a sign: "I'm sure I understand when I see the fingerspelling." At the lower grade levels, students responded positively to fingerspelling with signing.

Two elementary students commented that when used with the ASL sign, fingerspelling "was more helpful." Fingerspelling only was studied at the high school level and none of the 15 participating students reported preferring fingerspelling only. One high school student reported viewing the fingerspelling only version of an item twice because he was "thrown off with fingerspelling" and another student noted that the fingerspelling was fast and hard for him to see. These two comments provided evidence of some of the challenges that D/HH students face when test items contain fingerspelling only of key terms.

The third theme that emerged from the ASL cognitive lab data is related to item structure (diamond). Students preferred items signed in diamond structure, stating that the format was more like ASL than English. Two middle school students identified this contrast in their comments. One said the item set up in the non-diamond structure was "harder to understand" and the item with diamond structure "was better." The other student, who had less experience using ASL, echoed that statement by noting the item with diamond structure was "better" and used "more clear communication." This student further explained "ASL is beautiful; it shows more things" and the non-diamond structure "feels funny" and "doesn't make sense." High school students made similar comments. One student said the item presented in the diamond format was "more clearly explained" and the non-diamond structure was "confusing" and "tough."

Plurality and use of space, two issues studied at grades 3–5, yielded little information about students' preferences for different representations. In some instances, students were not able to discern a difference in how the two versions of an item set were represented. In these cases, the researcher explained how the items differed and probed students for information about whether one version was easier to access or more preferable. For plurality and use of space, no themes emerged from the data collected.

Randomized Controlled Trial

A total of 279 students from 17 states participated in the RCTs. All of these students were D/HH who normally use ASL support during assessment. Table 3 shows the distribution of participants based on teacher-rated reading and mathematics ability. Student ability was included in the design to ensure a comparable mix of students was distributed across the three test forms. The researchers opted to use teacher ratings of students' mathematics and reading ability instead of using a pre-test to keep the amount of testing time required for participation in the study to a minimum.

The purpose of the RCT was to examine the impact of embedded video ASL support and different versions of ASL videos on students' performance. Specifically, the study was designed to answer two key research questions:

1. Do D/HH students who normally receive ASL support during assessment differ in their performance on items delivered in a computer-based testing system with ASL videos and items without ASL videos?
2. Do D/HH students who normally receive ASL support during assessment differ in their performance on items with ASL Version 1 and ASL Version 2?

Table 3. Distribution of RCT participants by grade and teacher-rated ability

	Grade band			
	3–5	6–8	9–12	Total
Low reading and low math ability	20	68	92	180
Low reading and average or above math ability	18	16	27	61
Average or above reading ability and low math ability	4	8	2	14
Average or above reading and average or above math ability	12	7	5	24
Total	54	99	126	279

In order to answer the first research question, a paired sample *t*-test was employed. By administering the same 18 items, in the same order, with different versions of ASL support depending on form assignment, researchers were able to compare students' scores on the 12 items that were administered with support to the 6 items that were administered unsupported. Due to the difference in the number of supported and unsupported items on the test, the unsupported score was weighted for this analysis. The researchers hypothesized that students who normally receive ASL support for assessment would perform better on supported items, therefore, a one-tailed test was employed. In aggregate, students across the three grade-level bands, who normally receive ASL support during assessment, achieved a significantly higher mean score on the supported items ($M = 3.24$, $SD = 2.04$) than on the unsupported items ($M = 2.91$, $SD = 2.28$, $t\ (278) = 2.24$, $p = .013$). In order to provide evidence that the difference in supported versus unsupported scores can be attributed to the supports and not to differences in student ability, a chi-square analysis was conducted and confirmed that random assignment of students to test forms yielded an even distribution of abilities (based on teacher judgment of students' math and reading ability) across forms (Grades 3–5: $X^2(6, N = 54) = 5.43$, $p = .49$; Grades 6–8: $X^2(6, N = 99) = 7.38$, $p = .29$; Grades 9–12: $X^2(6, N = 126) = 2.45$, $p = .87$).

In order to answer the second research question, paired sample *t*-tests were used to compare students' scores on the three items that were administered with Support Variation 1 to the three items that were administered with Support Variation 2 for each issue. For each issue, students were administered six items, four of which were supported and two of which were not supported. The four supported items contained two different versions of ASL support for a given issue. For example, among items designed to study fingerspelling, math terms were fingerspelled only in support Variation 1 and math terms were signed and fingerspelled in Support Variation 2. For all variations studied, there were no statistically significant differences in performance between the two supported ASL versions (see Appendix for paired-samples *t*-test results).

Following the completion of the research phases of the project, the GAAP ASL team reviewed the draft guidelines used to create the ASL videos for research purposes, sought public input from schools for the deaf and state/consortia ASL experts, and engaged in consensus-driven discussions in light of the cognitive lab and RCT findings. This step was particularly important given that no statistically significant differences were found in performance between the two ASL versions of test items for all issues studied. The research findings combined with expert judgment led to decisions about appropriate strategies for representing content in ASL that measure the same construct as the English text version of the item, and associated guidelines were documented. The research team then identified items from the research test forms that illustrate application of the guidelines for use as exemplars.

Discussion

Lessons Learned

During the 2-year GAAP project, the team learned three key lessons about creating ASL test items. These lessons, which provided the foundation for the GAAP ASL Guidelines, are documented and explained in detail in the guideline document (Supplementary Materials). The first lesson is that it is critical to use a team-based approach to translation. When creating the ASL videos of test items, GAAP adopted a team structure where a deaf content expert who is also an educator with native ASL fluency and an English–ASL bilingual specialist who is also an experienced interpreter co-led the development of the ASL representations of item content. The goal of this duo was to recreate English items in ASL by adhering to the linguistic rules and conventions of ASL while not altering the constructs measured by the English-based items. The pair was supported by a content specialist with experience in the content area being assessed, an assessment accessibility specialist with expertise in measurement, video production specialists with experience in creating ASL videos, an ASL linguist to assist with translation issues, and interpreters to facilitate group discussions. Each team member brought important expertise that is essential in developing high-quality ASL videos.

The second lesson learned is that a multi-step process should be used in creating ASL test items. This team-based collaborative process should include a review of English test items and discussion of potentially controversial or otherwise challenging interpretation issues; development and review of draft ASL videos of items identified as potentially challenging or controversial; recording of ASL videos for all items; review of videos by external ASL and content experts. Ongoing research and evaluation is also recommended in order to make improvements to subsequent ASL videos of assessment content.

Lastly, the ASL videos developed using the GAAP guidelines produce high-quality ASL access that has a positive impact on test performance when compared with no ASL access, as evidenced by the RCT results. This is an important finding as no other research study has shown a positive impact of ASL support on students' test performance. Although no differences were found in the two ASL versions used in the RCT for each issue studied, key findings from the cognitive lab research shed light on issues related to translating test content into ASL. Findings from the cognitive labs support the idea that the translation should adhere to the linguistic rules and conventions of the language into which the items are being translated. Specifically, findings from the cognitive labs suggest that D/HH students are better able to understand items that (a) use ASL conventions related to the order in which information is presented (e.g., the diamond structure), (b) are consistent with how ASL is used during instruction (e.g., nontextual content such as equations and graphs is signed), and (c) are consistent with ASL conventions related to the use of fingerspelling. The appropriateness of fingerspelling terms in an assessment is a particularly important issue as several state and consortia guidelines recommend fingerspelling key terms, especially math and science terms, due to concern that the ASL sign provides too much construct-relevant information to the student. Based on evidence from GAAP research and secondary sources, combined with expert opinion, the GAAP team recommends limiting fingerspelling to cases where most students are unlikely to be familiar with an ASL term and where fingerspelling a term would be linguistically appropriate (e.g., lexicalized fingerspelling, neutral fingerspelling of proper nouns, abbreviations, two-word compounds, and signed-fingerspelled compounds). And, consistent with ASL convention and supported by findings from the GAAP cognitive labs, the guidelines suggest that when there is an ASL term available but students are unlikely to be familiar with it, the term should be signed, followed by the fingerspelled English equivalent.

Limitations

There are several limitations to the GAAP project that affect the generalizability of these findings. The first limitation is that item statistics were not available for the items used in the research. The GAAP team felt that it was important to use test items from the newly created pool of released college and career ready aligned test content created by states and consortia. However, at the time the GAAP researchers selected items from the pool of released items, only math and English Language Arts items were available and none of the items had been field-tested and therefore, item statistics were not available. As a result, we do not know if the items used in the RCT forms studying a particular issue are comparable in terms of item difficulty. Ideally, the six items used to study an issue, for example fingerspelling, would have similar item difficulties so that an individual students' score on the three versions of the fingerspelling items (no support, fingerspelling and sign support, and sign support only) would more likely be a function of the support variation rather than a function of item difficulty. Random assignment of students to forms lessened the impact of this limitation. A chi-square test verified that random assignment of students to test forms yielded an even distribution of abilities (based on teacher ratings of students' math and reading ability) across forms.

A second limitation is a lack of background information for students participating in the RCT. Information such as whether the student has other deaf family members, the age the student started learning ASL, the amount of time the student has received instruction in ASL, and a measure of ASL fluency was not collected nor used in the analysis of differences in students' performance on supported and unsupported items.

The third limitation is that the RCT included a small number of test items in only one content area (mathematics). The researchers capped the number of items on the RCT test forms to 18 plus one practice item to ensure that students could complete all of the items in a single testing session and to reduce student fatigue that often sets in toward the end of lengthy assessments. As a result, the number of supported and unsupported items on each RCT test form was not equal which required researchers to weight the unsupported score for some of the analyses. This is a limitation because it assumes that had there been additional unsupported items, students' performance would have been comparable. As previously noted, mathematics was chosen as a focus content area because state policies often do not allow any language translation for English Language Arts assessments. However, the GAAP Guidelines are intended to be used across state assessment content areas, because with only one exception (presentation of mathematical notation) the areas studied (fingerspelling, diamond item structure, use of space, and plurality) apply to test items in mathematics, reading, science, and social studies.

The last limitation is that non-ASL fluent researchers were used to conduct the cognitive labs, requiring an interpreter to facilitate communication between the student and researchers. The interpretation step may have impacted the accuracy of information documented during the cognitive lab sessions.

Conclusion

In comparison with the traditional ASL "on the fly" accommodation, video-embedded ASL requires significantly more upfront time and expense to develop, but the end result is far greater standardization and quality administration. Although computer-based delivery of assessments with ASL videos has the potential to dramatically improve both access to test content and measurement of proficiency, the success of this approach is dependent on the ability of test developers to appropriately and linguistically represent test content in ASL. The research presented in this article lays the foundation for further research to more deeply understand students' preferences for and the impact of different ASL translation decisions on students' performance on academic assessments. This future research should take into account student background data such as level of ASL fluency and home factors such as whether the student has deaf family members; utilize test items with known item characteristics in order to better understand differences in students' performance when translation variations are used; utilize a sufficient number of items per translation variation under investigation to detect differences; and strive for sample sizes with adequate power to detect differences between translation variations. More research is also needed to better understand students' interactions and preferences for accessibility features of computer-based test delivery systems. For GAAP, ASL videos were embedded in the test delivery system, allowing students to view the English text and ASL video of each test item simultaneously. Other ASL research and development projects have provided dual access to ASL and English by allowing students to toggle between full screen ASL video and full screen English text (Hoffmeister et al., 2013). Research on the presentation of videos and other features of test delivery systems could allow for a design that can provide students with high-quality access to assessment content.

Providing high-quality access to test content is critically important to the success of state tests aimed at measuring academic proficiency. The GAAP consensus- and evidence-based guidelines for the development of ASL versions of academic test content for K-12 students provide states with recommended qualifications of members of ASL item development teams, a recommended development process, ASL grammar guidelines, test content guidelines, item feature considerations, and filming considerations (Supplementary Materials). The guideline document along with ASL videos of the guidelines and item examples are an important tool that can be used by test developers to improve accessibility for D/HH students who use an ASL accommodation.

Supplementary Data

Supplementary material is available at http://jdsde.oxfordjournals.org/

Conflicts of Interest

No conflicts of interest were reported.

Funding

U.S. Department of Education Enhanced Assessment Grant Program (S368A120006). However, 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.

References

Ansell, E., & Pagliaro, C. M. (2001). Effects of a signed translation on the type and difficulty of arithmetic story problems. *Focus on Learning Problems in Mathematics*, 23, 41–69.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cawthon, S. W., Winton, S. M., Garberoglio, C. L., & Gobble, M. E. (2011). The effects of American Sign Language as an assessment accommodation for students who are deaf or hard of hearing. *Journal of Deaf Studies and Deaf Education*, 16, 198–211. doi:10.1093/deafed/enq053

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., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English language arts assessment of elementary, middle, and high school students*. Iowa City, IA: Pearson Education.

Ericsson, K. A., & Simon, H. A. (1999). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Massachusetts Institute of Technology.

Every Student Succeeds Act. (2015). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35. doi:10.1111/j.1745-3992.2006.00076.x

Hoffmeister, R., Reis, J., Henner, J., Beckert, B., Lopez, R., & Ferro, D. (2013). *Developing an online ASL STEM lexical database: The American Sign Language Vocabulary Reference Tool (ASL-VRT)*. Paper presented at The Convention of American Instructors of the Deaf, Rochester, NY.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Marschark, M., Pelz, J., Convertino, C., Sapere, P., Arndt, M., & Seewagen, R. (2005). Classroom interpreting and visual information processing in mainstream education for deaf students: Live or memorex®? *American Educational Research Journal*, 42, 727–761. doi:10.3102/00028312042004727

Marschark, M., Sapere, P., Convertino, C., & Seewagen, R. (2005). Access to postsecondary education through sign language interpreting. *Journal of Deaf Studies and Deaf Education*, 10, 38–50. doi:10.1093/deafed/eni002

Marschark, M., Sapere, P., Convertino, C., Seewagen, R., & Maltzen H. (2004). Comprehension of sign language interpreting: Deciphering a complex task situation. *Sign Language Studies*, 4, 345–368. doi:10.1353/sls.2004.0018

Maihoff, N. A., Bosso, E., Zhang, L., Fischgrund, J., Schulz, J., Carlson, J., & Carlson J. E. (2000). *The effects of administering an ASL signed standardized test via DVD player/television and by paper-and-pencil: A pilot study*. Dover, DE: Delaware Department of Education.

Qi, S., & Mitchell, R. E. (2012). Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future. *Journal of Deaf Studies and Deaf Education*, 17, 1–18. doi:10.1093/deafed/enr028

Russell, M., Kavanaugh, M., Masters, J., Higgins, J., & Hoffmann, T. (2009) Computer-based signing accommodations: Comparing a recorded human with an avatar. *Journal of Applied Testing and Technology*, 10. Retrieved from http://www.testpublishers.org/assets/documents/computer_based.pdf

Schick, B., Williams, K., & Kupermintz, H. (2006). Look who's being left behind: Educational interpreters and access to education for deaf and hard-of-hearing students. *Journal of Deaf Studies and Deaf Education*, 11, 3–20. doi:10.1093/deafed/enj007

Texas Education Agency. (2012). General Instructions for Administering Statewide Assessments to Students Who are Deaf or Hard of Hearing.

Thurlow, M., & Kopriva, R. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education*, 39, 331–369.

Thurlow, M. L., & 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., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements*. Thousand Oaks, CA: Corwin Press.

Willis, G. B. (1999). Cognitive interviewing: A "how to" guide. Paper presented at the Meeting of the American Statistical Association, Research Triangle Institute, North Carolina, NC.

Appendix. Comparison of students' performance on items with Support Variation 1 and items with Support Variation 2

		Paired samples test							
		Paired differences							
					95% Confidence interval of the difference				
		Mean	*SD*	*SEM*	Lower	Upper	*T*	*df*	Significane (two-tailed)
G3-5 Pair 1	Fingerspelling (fingerspell vs. fingerspell+sign)	−.15	1.02	.14	−.43	.13	−1.07	53	.289
G3-5 Pair 2	Use of Space (general vs. specific)	.19	1.10	.15	−.12	.49	1.24	53	.222
G3-5 Pair 3	Plurality (plural vs. singular)	−.02	.86	.12	−.25	.22	−.16	53	.875
G6-8 Pair 1	Fingerspelling (fingerspell vs. sign)	−.01	.90	.09	−.19	.17	−.11	98	.911
G6-8 Pair 2	Diamond (diamond vs. no diamond)	−.01	.81	.08	−.17	.15	−.12	98	.902
G6-8 Pair 3	Equations (equations sign vs. no sign)	−.07	.85	.09	−.24	.10	−.83	98	.409
G9-12 Pair 1	Fingerspelling (fingerspell vs. sign)	−.09	.92	.08	−.25	.08	−1.1	125	.289
G9-12 Pair 2	Diamond (diamond vs. no diamond)	−.01	.84	.08	−.16	.14	−.11	125	.916
G9-12 Pair 3	Equations (equations sign vs. no sign)	.02	.77	.07	−.11	.16	.34	125	.731