Challenges and solutions in test adaption

Comparing international experiences with the *British Sign Language Production Test (Narrative Skills)*

Charlotte Enns[1], Vera Kolbe[2] and Claudia Becker[2]
[1] University of Manitoba | [2] Humboldt-Universität zu Berlin

Sign language assessment tools are important for professionals working with DHH children to measure sign language development and competence. Adaptation of an existing test can be a solution when initiating assessment in a sign language community; the adaptation process must adhere to key principles and procedures. We introduce the principles of test adaptation and outline the challenges we faced in adapting the *British Sign Language Production Test* (Herman, Grove, Holmes, Morgan, Sutherland & Woll 2004) to German Sign Language and American Sign Language. Challenges included decisions regarding the normative sample, the use of terminology, and variations in the scoring protocols to fit with each language. The steps taken throughout the test adaptation process are described, together with a comparison of parallels and differences. We conclude that test adaptation is an effective method of developing practical tools for sign language assessment and contributes to a better understanding of sign language development.

Keywords: sign language assessment, test adaptation, test adaptation principles, test adaptation challenges, score sheets, sign language development

1. Introduction

Sign language assessment serves a variety of purposes in the education of deaf and hard of hearing (DHH) students. One major purpose of assessment is determining children's level of sign language proficiency to track typical development and make decisions regarding intervention and education. A second purpose is to monitor children's progress as they proceed through their education. Areas in

Videos available from https://doi.org/10.1075/sll.20010.enn.video
https://doi.org/10.1075/sll.20010.enn | Published online: 19 July 2021
Sign Language & Linguistics 24:2 (2021), pp. 226–258. ISSN 1387-9316 | E-ISSN 1569-996X

which DHH children are having difficulties developing sign language skills are often identified by professionals through assessment. Identification of acquisition difficulties and strengths is a third purpose of assessment, e.g., to provide important guidelines for language teaching. Finally, assessment is also required for reporting purposes so that parents are aware of their child's level of competence and rate of progress. Accurate assessment can serve a variety of purposes and indicates the need for effective sign language assessment tools. Although these tools are needed, to date very few assessment measures for sign language acquisition are available, primarily only for American Sign Language (ASL) and British Sign Language (BSL) (Enns, Haug, Herman, Hoffmeister, Mann & McQuarrie 2016). As a result, many educational programs and teachers have to rely on informal descriptive measures to develop teaching goals and monitor progress (Herman 1998, 2015; Kolbe 2019; Singleton & Supalla 2011).

The purpose of this article is to share our experiences of adapting an existing sign language test for use in different sign languages. In the introduction, we provide a brief overview of test development options and our decision to adapt an existing test from another sign language. In Section 2, we expand on the principles of test adaptation that were used throughout the process. Section 3 provides a description of the original test, and in Section 4, we describe the structure of the test adaptation process following the guidelines of the International Test Commission (2017) and provide examples of equivalence levels as introduced by Iliescu (2017). Section 5 addresses the specific challenges we encountered and the ways that we resolved these issues. We conclude with some suggestions for future research. The process we outline here may serve as a guide for other researchers considering test adaptation to increase the availability of sign language tests for a variety of different sign languages.

One strategy to increase available sign language assessment measures is to adapt existing tests for use in other sign languages. Key examples of implementing this strategy are the various sign language versions of the *British Sign Language Receptive Skills Test (BSL RST)* (Herman, Holmes & Woll 1999), which was the first standardized test of any sign language in the world that was tested for reliability and validity (Johnston 2004). Reliability and validity are key psychometric characteristics used to describe the quality of a test. Language tests measure performance at a given time with a given task, therefore the quantification of language competence involves descriptive information of a complex construct (Jones 2012). Results from language tests are used to compare, describe and document language competence in a specific area. Reliability is the consistency of the test, the extent to which the same rank order is replicated. Validity refers to the accuracy of the test in measuring what it is intended to measure (Council of Europe 2001). In their socio-cognitive framework, O'Sullivan & Weir (2011) use social,

cognitive and evaluative aspects to define validity, thus also focusing on the social consequences of test use. Validity is described as construct validity with multiple interactive layers. One of the layers is criterion-related validity, including reliability. Steps to ascertain validity are further described in Section 4 outlining the test adaptation processes. An additional criterion of the quality of a language test is feasibility (Council of Europe 2001), meaning that the test can be taken and assessed in a limited amount of time. In the case of a test targeting children, child-appropriateness is another criterion the test needs to meet.

The *BSL RST*, as a standardized test, meets the demands of reliability, validity, feasibility, and child-appropriateness, which is why researchers from several different countries have adapted the *BSL RST* into other sign languages, including ASL (Enns, Zimmer, Boudreault, Rabu & Broszeit 2013), Spanish Sign Language (Valmaseda, Pérez, Herman, Ramírez & Montero 2013), Italian Sign Language (Meristo, Falkman, Hjelmquist, Tedoldi, Surian & Siegal 2007), German Sign Language (DGS) (Haug 2011), and Australian Sign Language (Johnston 2004). Another test originally developed in BSL, the *British Sign Language Production Test (BSL PT)* (Herman, Grove, Holmes, Morgan, Sutherland & Woll 2004) is also in the process of being adapted for use in ASL (Enns, Zimmer, Broszeit & Rabu 2019) and DGS (Kolbe 2019), and these adaptation experiences form the basis for the current article.

Another strategy for developing sign language tests is the adaptation of an existing test from the societal spoken language. This strategy benefits from cultural similarities (Novogrodsky & Meir 2020: 818) but may require considerable adaptation for the linguistic differences, while the adaptation from a sign language test benefits from modality-specific linguistic similarities but requires adjustments for cultural differences. There are certainly formal sign language tests that have not been adapted from either signed or spoken languages, but rather have been created specifically to assess aspects of a sign language following culturally appropriate development processes, such as involving deaf professionals and native signers. An example of such a test is the American Sign Language Assessment Instrument (ASLAI) (Hoffmeister, Caldwell-Harris, Henner, Benedict, Fish, Rosenberg, Conlin-Luippold & Novodgrodsky 2014). The ASLAI is modeled on tests for spoken language development and tests of reading achievement, measuring conversational abilities, academic language knowledge, language comprehension, analogical reasoning, and metalinguistic skills. The ASLAI is designed to test DHH students between the ages of 4 years and 18 years. There are 12 subtasks in the total battery, divided into four categories: (a) tests of vocabulary (including Simple Vocabulary, Antonyms, Synonyms, and Vocabulary in Sentences), (b) tests of reasoning skills (Analogies – measuring relationships based on causal, purpose, antonym, noun-verb pairs, and phonology), (c) tests of syntax (including Classifier Sorting Task, Real Objects and Plurals – measur-

ing Verbs of Motion and Verbs of Location), and (d) one test of ASL text comprehension (multiple choice comprehension questions to extract both literal and inferential meaning from ASL texts). Importantly, the ASLAI provides a measure of the relationship between specific areas of children's receptive signed language abilities and the comparable, and critical, areas of their English literacy skills.

There is a need to increase available sign language assessments, whether through unique development or adaptation of existing spoken or signed language tests. Each of these strategies has advantages and disadvantages, but here we will focus on how we, as researchers from different countries, independently decided to adapt an existing sign language test. We will discuss the procedures and challenges encountered, as well as lessons learned, through our experiences of adapting the *BSL PT*.

2. Principles of test adaption

The process of adapting existing sign language tests into other languages was examined by Haug & Mann (2008). They clarified an important distinction between "translation", defined as a one-to-one transfer without consideration of linguistic differences, and "adaptation", which involves developing a parallel test that "acknowledges the linguistic, cultural, and social conditions of those taking the adapted test while retaining the measurement of the constructs found in the original" (Oakland & Lane 2004: 239). Consideration of linguistic differences includes the recognition of how depicting verbs fall into different categories in BSL, ASL and DGS. Cultural differences can be as simple as the shape and color of a British vs. German mailbox or as complex as appropriate representation of diversity in story characters.

In recent years, the interest in research on test adaptation in cultural and linguistic contexts has evolved internationally. One of the major contributors is Iliescu (2017) who gives a very profound overview of different international guidelines concerning test adaptation, describes critical phases in the test adaptation process, and provides checklists to prevent construct bias, method bias and item bias in the different phases of test adaptation. Another important publication describing the phases of the test adaptation process is the "Guidelines for Translating and Adapting Tests" published on the homepage of the International Test Commission (ITC 2017).

Iliescu (2017) describes test adaptation as a scientific process that is guided by the principles of the scientific method and needs to "offer proof for the appropriateness of the […] linguistic transformation not only in terms of language, but also in terms of psychometric characteristics" (Iliescu 2017: 19). In the mostly iter-

ative process of test adaptation, it is important to strive for equivalence as a special form of validity when comparing test scores across different groups and testing processes. The process "is associated with measurement aspects and interpretive aspects of cross-cultural comparison" (Iliescu 2017: 131). Equivalence needs to be provided on different levels, at the beginning as linguistic equivalence, providing the smallest possible invariance of the initial test translation, and later as psychological equivalence encompassing cultural and psychometric equivalence. The different levels of equivalence are defined as the lack of bias related to those levels. Construct bias describes an "incomplete overlap of the measured constructs in the original and adapted versions of the test" (2017: 140), e.g., checking for cultural and linguistic appropriateness. Method bias is described as "nuisance factors arising from aspects of method" (2017: 140), e.g., ensuring comparability of video stimuli. Item bias implies "anomalies in items" (2017: 140), e.g., revising items that do not occur consistently. We elaborate on these levels of equivalence in Section 4. If the goal is to develop a test that closely resembles the existing test, but incorporates the specific needs of the target language, which was definitely the case with developing the *BSL PT* for use in ASL and DGS, then adaptation is the appropriate term to use to describe the process.

The advantage of adapting an existing test rather than developing an original test is that important test design considerations and decisions have already been evaluated. These considerations include the selection of grammatical features that are important indicators of proficiency, the composition of normative samples, and valid assessment task formats. For example, the *BSL PT* is based on what is known about sign language acquisition and highlights grammatical features identified in the research as important indicators of proficiency, such as verb morphology (e.g., spatial verbs) and use of space (e.g., role shift) (Herman et al. 1999). Considering that many sign languages share these important grammatical features, it is likely that test items will be relevant in sign languages other than BSL (Neidle et al. 2000; Schick 2010). Iliescu (2017: 55) describes practitioner's need as one of the forces behind test adaptations, definitely a need that is felt by teachers and language therapists working with DHH children. Economic reasons may also play a role; although test adaptation is a costly undertaking, lessons can be learned from the original test development and mistakes avoided.

Sample size is a problematic factor in our field of study, since our selection criteria and the smallness of the Deaf community limit the possible number of participants (Johnston 2004; Schembri et al. 2002). If the study focuses on only native-signing deaf children, the number of potential participants is reduced even more due to the low incidence of deaf children being born to deaf parents. Low incidence can also require researchers to travel to various locations within a country to collect appropriate data, which may be costly or not feasible. There-

fore, the implementation of test adaptation provides the benefit of cross-linguistic comparison, where the results in the original test language can support or contradict the findings in the new language.

Another important consideration related to sample size is the composition of the standardization sample, given the inconsistent exposure to sign language that occurs for most DHH children (Hall 2017; Mitchell & Karchmer 2004). Decisions regarding the inclusion of hearing children with Deaf parents, or DHH children with hearing parents attending bilingual or Total Communication/Sign Supported Speech programs have already been made and substantiated with research evidence for the *BSL PT*. In addition, clear guidelines for the assessment format have also been validated. These decisions included: using a language-free video with a story that is engaging for children; focusing on grammar and narrative abilities rather than specific vocabulary; keeping the video story to an appropriate length to avoid excessive memory load; using a story plot with repetitive sequences and a vivid climax; and incorporating mandatory training on the test scoring system to minimize influence by the test administrator.

In order to produce a culturally and linguistically sensitive test adaptation, it is imperative to follow the demands of the Deaf community regarding ethical considerations implied in the underlying conduct of research. These demands can be compared to demands for culturally sensitive research ethics voiced by diverse linguistic and cultural minorities (Harris, Holmes & Mertens 2009: 112). Appropriate ethical conduct for research in sign language communities was an important consideration in conducting the background studies of the test adaptation process. The Ethics Statement for Sign Language Research of the Sign Language Linguistic Society (2016) indicates three areas that need to be considered: responsibility to Deaf individuals, responsibility to Deaf communities, and responsibility to scholarship and to the public. Efforts need to be made to involve members of the Deaf community and – if possible – deaf researchers in all stages of the adaptation process. The Deaf community should be informed about the research on their sign languages, e.g., on a free accessible homepage. Special care must be taken during the research with children to ensure their well-being and foster their pride in their own language competence.

The decision of whether it is advantageous to adapt an existing instrument that has already been tested and standardized must be considered within the framework of evaluating the linguistic and cultural differences between the original and target languages. The current discussion works within such a framework and, therefore, provides valuable insights into the similarities and differences between assessing the narrative abilities of children learning BSL, ASL and DGS. Some of these differences were easily resolved through the modification of test stimuli, but others required more significant changes to the test scoring and

analysis system. The study also reinforces the benefit of collaboration among researchers in advancing a better understanding of natural sign language acquisition and measurement.

3. Description of the original test

The *BSL PT* (Herman et al. 2004) is a narration task based on children watching a three-minute language-free video presented on a TV/computer screen. The *BSL PT* provides mean scores for children 4–11 years old. The video, "Spider Story", features a boy and a girl acting out a series of events without communicating to each other in either signed or spoken language. The video plot is a sequence of similar events that are easy to remember and require different grammatical strategies while telling the story. The thrilling climax, when the boy is biting into a spider hidden in a sandwich, is impressive, so it is even recalled and narrated by very young children. Children are told that they will watch a video and then tell the story to a deaf BSL user who has not seen the video. After narrating the story, the child answers three pre-recorded questions targeting story comprehension and inferencing skills. The child's story (narrative) and responses to questions are video recorded for later analysis.

Scoring of samples is based on coding three areas: (i) Narrative content (16 narrative episodes, responses to questions); (ii) Narrative structure (orientation, complicating actions, climax, resolution, evaluation, and sequence); (iii) BSL grammar (spatial verbs/depicting verbs, agreement verbs, manner inflections, aspectual inflections, and role shift). The design of the original score form is well-suited for practitioners, in that it provides clear instructions of what to score and in which linguistic category items belong. Figure 1 provides an example of the *BSL PT* score form.

Narrative content *Score child's spontaneous story, without prompts*		**Narrative structure**		**BSL grammar** *Include child's spontaneous story and responses to questions*	
• girl brings tray with food (may include description of food on tray)	1	**Orientation** (setting the scene) *Score 1 for 1 reference to orientation*	1	**Spatial verbs** (GIRL) BRING-TRAY • correct handling classifier for TRAY • movement of BRING from one location to another, as if carrying the tray OR putting it down on a surface	1
• boy watches tv	1	*Score 2 for 2 references*	1	PERSON-GO/WALK-TO/FROM (TRAY) • correct person OR legs classifier • movement ot GO/WALK towards OR away from location of tray	1
• girl takes sweet, unwraps, about to eat sweet, boy demands sweet, girl gives sweet	1	**Complicating actions 1** *(events preceding the climax)*		X-PICK-UP-X • correct handling classifier for SWEET or CAKE or CUP • movement of PICK-UP away from location of tray	1
• girl takes cake, unwraps, about to eat	1	*Score 1 for 1 reference to complicating actions*	1		

Figure 1. Example from *BSL PT* score sheet (Herman et al. 2004)

Following analysis of a child's story and responses to questions, the raw scores obtained can be converted to percentiles based on five age groups. It is also possible to analyze a child's performance according to the narrative and grammatical features tested to identify strengths and weaknesses and identify targets for intervention.

The format of the *BSL PT*, being a narration task based on a language-free video, has good potential for adaptation into other sign languages. The BSL grammatical structures assessed (spatial verbs, agreement verbs, aspect, manner, and role shift) also fit well with both ASL and DGS grammatical categories. As a result, adaptation to the specific features of how these grammatical structures are marked in ASL and DGS was considered to be somewhat straightforward. Our test adaptation processes were structured into several phases, resulting in similarities and differences as outlined in the following section.

4. Structure of test adaptation process

The guidelines described by the International Test Commission (2017) provided the structure for our test adaptation phases, including Pre-conditions, Test Development, Confirmation, Administration, Score Scales, Interpretation, and Documentation.

Table 1. Steps in test adaptation process: I. Pre-conditions

I. PRE-CONDITIONS	
EXPERT INTERVIEW with ASL/DGS-linguists and Deaf researchers concerning linguistic categories	construct equivalence
LITERATURE RESEARCH concerning linguistic categories, sign language acquisition, research ethics in Deaf communities	construct equivalence
PERMISSION OF ORIGINAL TEST HOLDERS research team of BSL Production Test, participation in training course	

The Pre-conditions phase is necessary to determine if there is enough overlap in definitions and content of the construct for the intended use in different populations (ITC 2017: 9). An effort has to be made to minimize cultural and linguistic differences (ITC 2017: 10). As indicated in Table 1, both adaptations started with expert interviews with ASL/DGS linguists and Deaf teachers or Deaf researchers to determine the feasibility of test adaptation. Experts were asked whether the

linguistic categories (spatial verbs, agreement verbs, aspect, manner, and role shift) are also identifiable in the respective sign language. For example, "Will the video elicit a narrative that includes the grammatical target structures?" or "Do we have aspectual modifications in DGS/ASL? Can you give some examples?" In all cases, the answers from experts were positive and supported the transferability of the original grammatical structures from the *BSL PT*.

In addition to the support from sign language experts, literature reviews were conducted to check whether background research on the intended linguistic categories was already available in ASL and DGS. The aim was to confirm whether similar grammatical categories have been identified and researched in the target sign languages, and whether acquisition studies of these structures are available. For DGS there is very limited research regarding sign language acquisition. As previously mentioned, one of the benefits of test adaptation is that decisions regarding selection of grammatical features in the test have already been made. Thus, test adaptation provided the possibility of producing a test for DGS even though acquisition studies are mostly lacking. Hänel (2005) studied DGS acquisition of verb agreement with two children longitudinally (aged 2;2–3;3/3;4). Becker (2009, 2018) studied DGS acquisition of narrative competences in four studies. Specifically, the literature reviewed related to DGS included:

- Narrative competences in DGS (Becker 2009; Becker, Hansen & Barbeito Rey-Geissler 2018)
- Aspect and manner in DGS (Happ & Vorkörper 2014; Schwager 2012)
- Verbs in DGS (Papaspyrou, Meyenn, Matthaei & Herrmann 2008; Erlenkamp 2012; Hänel 2005)
- Constructed action (Fischer & Kollien 2006)

For ASL far more acquisition studies are available, so the support for the linguistic categories included in the test was based on literature reviews in the following areas:

- Agreement verbs (Meier 2002)
- Directional verbs (Lillo-Martin & Meier 2011)
- Spatial verbs (Anderson 2006; Meier 1991; Schick 2002)
- Aspect and manner (Simms, Baker & Clark 2013; Singleton & Newport 2004)
- Role shift (Cokely & Baker 1991; Klima & Bellugi 1979)
- Narrative abilities (Cravens 2013; Emmorey & Reilly 1998; Reilly 2005; Singleton & Morgan 2006)
- Various grammatical features included in review articles (Chen Pichler 2012; Haug 2011; Schick 2010)

For the DGS adaptation, ethical guidelines for research with minority groups, especially Deaf communities were collected. As is necessary at this stage, both research teams contacted the original BSL Production Test research team: Rosalind Herman, Nicola Grove, Sallie Holmes, Gary Morgan, Hilary Sutherland, and Bencie Woll. This resulted in a very fruitful cooperation, as they are experienced with international adaptations of their tests, they provide their own guidelines for test adaptation and enabled the participation of the foreign researchers at their training workshop for future test users. The permission for the adaptation was granted by the original test holders.

Table 2. Steps in test adaptation process: II. Test Development

II. TEST DEVELOPMENT	
TEST DEVELOPMENT considering linguistic development, cultural suitability of test content and test instructions	construct and item equivalence
PRE-TEST with Deaf adults ($n = 2$ for DGS; $n = 15$ for ASL) resulted in item adaptation	item equivalence
PRODUCTION OF NEW INPUT VIDEO cultural appropriateness of video setting, character diversity, visibility of hearing technology	construct equivalence
PRODUCTION OF 3 PARALLEL VIDEOS – ONLY ASL ensuring equivalency of narrative content, structure and grammar	
INTERVIEW – ONLY DGS with deaf and hearing kindergarten children ($n = 5$) to ensure familiarity with the cultural setting used in the video	item equivalence
STRUCTURAL ANALYSIS OF VIDEOS comparability concerning scene length and use of space in the plot as sign language influencing features	method equivalence
PILOT-TESTING – ONLY ASL native signing deaf children ($n = 47$) to ensure equivalency between 3 parallel videos, correlation between age and score	
QUESTIONNAIRE – ONLY DGS regarding cultural appropriateness of input video with members of Deaf community and majority community ($n = 6$) resulting in	
COMPARATIVE STUDY to check for acclaimed distractors in new input video with hearing children ($n = 5$) and Deaf adults ($n = 2$)	method and item equivalence
TEST MATERIAL DESIGN excel score sheets: machine-readable (ASL), user-oriented (DGS)	

In the Test Development phase (see Table 2), the translation and adaptation process must consider linguistic, psychological and cultural differences (ITC 2017: 11), and evidence needs to be provided that test instructions and item content have similar meaning for all intended populations. Item formats, categories and modes of administration need to be checked for suitability in the intended populations. A pilot study is recommended (ITC 2017: 15). However, in the DGS adaptation process, a pilot study with Deaf native signing children was not conducted, since the number of possible children is low, and these children were needed for the quantitative study in the Confirmation phase (intended to standardize the adapted DGS Production Test). Therefore, a number of small studies was conducted. A pre-test using the original input video was conducted with Deaf adults and resulted in an item adaptation since one of the original items was not used by any of the pre-test participants in DGS. Although it is possible to produce a modification as assessed by the *BSL PT* manner item "(GIRL) HUNGRY/THIRSTY$_{\text{intensifier}}$" with an exaggerated movement and emphasizing facial expression in DGS, this was not produced by any of the Deaf adults. As a result, a substitute item was created for the DGS test (named NaKom DGS – *Narrative Kompetenzen in Deutscher Gebärdensprache*): "(GIRL) ANNOYED/ANGRY$_{\text{intensifier}}$". Further descriptions and videos of the two manner items are provided in Section 5.3, subsection "Item substitution". The original item and the new substitute item were both kept for scoring in the quantitative study and marked for subsequent evaluation.

A new input video for DGS was produced. The video plot and the idea of child characters representing diversity was maintained. As the BSL video is about 20 years old, some adaptation to current developments was necessary: for further identification possibilities, one of the characters was a Cochlear Implant user, which is also visible in the video. As it was intended to substitute the television in the input video with a laptop computer, an interview with five signing and speaking kindergarten children was conducted to make sure they could name the activity of watching a movie on a laptop computer. The children were all able to name it, some called it a tablet, which is acceptable within the context of the test. To minimize possible influences of the new input video on sign language features, the use of space and manual activities was analyzed as well as the time lengths of the video cuts. The new input video was also culturally adapted. The background setting, clothing of characters and children playing the characters was adapted following the culture definition of Reckwitz (2011: 6). To check this adaptation for cultural appropriateness, a questionnaire was distributed to members of the Deaf community and members of the majority community ($n = 6$). The questionnaire was divided into items assessing aesthetic world interpretations and intellectual world interpretations based on the cultural definition of Reckwitz (2011: 6). The

aesthetic world interpretation was operationalized as the background setting of the film, e.g., furniture and room, the food eaten by the children, and the clothes. The intellectual world interpretation contained in the input movie was operationalized as the commonality of children watching a movie on a laptop and children being afraid of spiders.

Features of the video that were marked as critical in the questionnaire were checked for influence in a comparative study with hearing children ($n=5$) to assess influence regarding the narration and with two Deaf adults to check for distracting influences on DGS constructions. For example, one respondent noticed that the appearance and clothing of one of the characters is a little bit "atypical". Therefore, we checked whether the appearance and clothing could result in a diversion from the main plotline, such as evoking a long description of one character. Another respondent assumed that the presented food in the video clip is "unhealthy". We also checked whether this was a distractor from the story plot. However, no bias towards the criticized content was found, and, for the most part, the criticized aspects were not even mentioned in the narrations. The DGS team did not test this with DHH children, due to the necessity to involve all DHH children in the normative sample.

The initial steps in the Test Development phase for the ASL test adaptation were similar to the DGS process. The original input video (Spider Story) was used to elicit ASL narratives from 15 Deaf adults to determine the specific narrative and grammar features that would be included by competent ASL users. Appropriate revisions were made to the scoring and analysis procedures based on the data collected. For example, the specific ASL handshapes used in spatial verbs indicating "picking up" or "handing over" objects differed from BSL and were revised. In addition to producing an updated input video of the original "Spider Story", the ASL researchers created two additional videos to ensure culturally appropriate stimuli and allow for re-testing of students. The process included developing a template to confirm that parallel features occurred across all three stories, writing new stories, videotaping the enactment, pilot testing with native ASL users, and developing guidelines for story analysis (score forms). The reason for developing three different stories/versions of the test (parallel elicitation videos) was to prevent children from becoming overly familiar with one story through the process of longitudinal testing, or pre- and post-testing procedures for research purposes. The procedure of developing several comparable versions of tests is common for spoken language tests (e.g., MAIN, Gagarina et al. 2012 – versions for various languages; PPVT, Dunn & Dunn 1997 – Form A and B). In addition, it was an opportunity to improve the technical quality of the video, remove visual distractions (plain furniture, walls, dishes, etc.), and increase diversity and gender variations

of characters (for more details regarding the creation of the three video stories, please refer to Boudreault, Zimmer & Enns 2015).

As ASL has the advantage of being a sign language that is fairly widespread, signed not only in the USA, but also in areas of Canada, the ASL research team was able to conduct a pilot study of the adapted and new test versions on a sample ($n = 47$) of typically developing native signers within the recommended age range of 4–13 years to determine effectiveness and reliability of scoring guidelines. Changes were made to the scoring system based on the pilot testing results, specifically in the areas of Role Shift and Narrative Structure/Content. These changes are detailed in Section 5.3. In addition, the pilot testing was conducted to provide evidence that the three different versions (stories) were parallel and could be used for comparison across students and assessments. Both research teams changed the score sheet design from paper to an excel sheet to facilitate multiple levels of analysis.

Iliescu (2017) describes the critical phases of test adaptation as translation design, pre-testing and norming, which is reflected in the multitude of steps that were taken in these phases by the researchers. In the Confirmation phase (see Table 3), a major decision is choosing a sample with relevant characteristics and sufficient size for empirical analysis. Construct, method and item equivalence needs to be confirmed, and statistical analysis is used to check for validity, including reliability, of the adapted test and to develop standards for the intended population (ITC 2017: 18). A large quantitative study for standardization was conducted in both adaptations. Both tests included an inter-rater reliability study, and the German statistical analysis included an intra-rater reliability study as well.

The reliability of the ASL test was investigated using inter-scorer comparisons and test-retest analyses with the three versions of the test. Inter-scorer reliability was assessed by having 10% (30 videos) of the data independently scored by two different trained testers and comparing the results. Statistical analysis using Pearson's correlation resulted in a highly significant correlation of 0.87 ($p < 0.01$), indicating that inter-scorer reliability was very good. The test-retest reliability was based on the 47 children participating in the pilot testing, who each completed two versions of the test (retold two different stories) within the same testing session. Parametric statistics (ANOVA) were used to compare the children's scores between all combinations of the three stories, and no statistical evidence was found for any differences (Sig = 0.288, greater than 0.05). A Test of Homogeneity of Variances was used to validate the assumption of the homogeneity for ANOVA and had similar results (Sig = 0.234, greater than 0.05). These analyses indicate strong test-retest reliability within participants and across all three versions of the test.

Table 3. Steps in test adaptation process: III. Confirmation

III. CONFIRMATION	
SAMPLE DEFINITION	
DGS DHH children age 4–11, early exposure to DGS from birth or at least in kindergarten, no additional disabilities, confirmed with parent + teacher background questionnaire	
ASL DHH children age 4–13, early exposure by age 3 years, confirmed with parent + teacher background questionnaire, normal non-verbal IQ, confirmed with TONI-4	
QUANTITATIVE STUDY to ensure reliability and validity and develop national standards, create national scoring guidelines (DGS: $n=97$, ASL: $n=215$)	construct, method and item equivalence
INTER-RATER RELIABILITY STUDY assessment of about 10% of data by two independent annotators, at least one Deaf annotator	
INTRA-RATER RELIABILITY STUDY – ONLY **DGS** second assessment of 10% of the data by the same annotator	
TEST-RETEST-RELIABILITY – ONLY **ASL** parametric statistical comparison of the same child's results with two different video stories ($n=47$)	

The validity of the ASL test, checking if the test was measuring what it was designed to measure (children's ASL abilities), was also determined. Scores (based on the categories of average, above average, below average) from the ASL test were compared with the same children's scores on the *ASL Receptive Skills Test* (Enns et al. 2013) using a Pearson's correlation. A highly significant correlation (0.91, $p < 0.01$) was found, suggesting good concurrent test validity.

For research purposes, the benefits of video-analysis through ELAN (computer software, 2019) were observed during the inter-rater reliability study for the DGS adaptation. In this study, 10% of the narrations were evaluated by two independent evaluators, one of them a Deaf research assistant. The evaluators marked in ELAN on the child's re-telling video exactly where the point in a specific category was given. This enabled a very detailed discussion of evaluation similarities and differences. ELAN is an excellent tool for the detailed analysis required for research purposes; however, the complexity of administering this tool effectively does not make it practical for use in schools.

As there is no other DGS assessment tool available, validity was checked by answering the questions proposed by O'Sullivan & Weir (2011: 20–21) for the different layers of validity:

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? Is the test likely to be appropriate for the candidates? (Test-taker)
- Are the characteristics of the test tasks and their administration fair to the candidates who are taking them? (Context validity)
- Are the cognitive processes required to complete the tasks appropriate? Are candidates likely to use the same cognitive processes as they would if performing the task in a "real world" context? (Cognitive validity)
- To what extent can we depend on the scores on the test? What do the numbers or grades mean? (Scoring validity)
- What effects does the test have on its various stakeholders? (Consequential validity)
- What external evidence is there outside the test scores themselves that the test is doing a good job? (Criterion-related validity)

It is important to emphasize that criterion-related validity does include reliability. For NaKom DGS an inter-rater-reliability study as well as an intra-rater-reliability study was conducted. Due to financial limitations and the collection of data on many occasions all over Germany, a test-retest analysis was impossible. The NaKom DGS study consists of 103 narrations of children. For the inter-rater agreement analysis, 21 narrations (20%) were randomly selected and independently assessed by two different trained testers. As a measure for agreement Cohen's Kappa was estimated (Levshina 2015: 201) as well as the correlation coefficient Pearson's r (Levshina 2015: 116). Both measures suggest strong agreement, with Cohen's Kappa 0.78 and Pearson's r 0.98 ($p < 0.001$). To check for the internal rater consistency, 10 narrations (10%) were re-assessed by the same tester one year later. Statistical analysis suggests strong agreement, with Cohen's Kappa 0.80 and Pearson's r 0.99 ($p < 0.001$).

As all analyses provided good results, both teams decided to continue to the publication phase. Steps that have to be completed for publication (finalized for the ASL adaptation, and in process for the DGS adaptation) include:

- Final design and production of test materials;
- Developing training materials and procedures for implementing the test so that it is accessible to schools/teachers:
  - Who can administer the test?
  - Who can score/analyze the test?

- How to ensure consistent administration and scoring of the test?
- How is the test distributed/made accessible to educators and researchers?

5. Challenges and solutions from two adaptation processes

We would like to highlight several challenges of the test adaptation process, including the determination of normative samples, terminology for sign language grammatical structures, and revisions made to the scoring system.

5.1 Normative samples

As previously mentioned, one of the benefits of adapting an existing test is that decisions have already been made regarding the composition of the normative sample. We followed the example of the *BSL PT* and included only children with early exposure to sign language and whose nonverbal intelligence was within the normal range. The criteria for the selection of the children were the same in both research teams: normal intelligence, early exposure to ASL/DGS, ASL/DGS before school, ASL/DGS used in everyday communication, no other impairments that could influence language acquisition. Teacher and parent questionnaires were used to obtain this information. Yet the operationalization of these criteria differs in the two adaptation processes.

In the case of the ASL test, all recruitment for normative testing, for a total of 215 students, occurred through schools for the deaf implementing bilingual (ASL and English) programming (see Table 4 for details). It was challenging to consistently and reliably collect background information from parents, particularly regarding languages used at home. As a result, early exposure to ASL was confirmed through the children's attendance in the school's preschool program or having at least one Deaf parent. Although these parameters (preschool participation and a Deaf parent) do not guarantee early exposure to ASL and considerable variability may still occur across DHH children's ASL development (Henner, Caldwell-Harris, Novodgrodsky & Hoffmeister 2016), normative testing found no significant difference between the DHH children with hearing parents and those with Deaf parents. These results verified a similar (and early) age of exposure to ASL for all children included in the normative sample. In order to ensure that children had average nonverbal intelligence, the *Test of Nonverbal Intelligence* (TONI-4) (Brown, Sherbenou & Johnsen 2010) was administered to most children, unless they were too young (less than 5 years of age) or the school had conducted similar testing within six months. The ASL team recognizes that controlling

for early exposure to Sign Language (ASL) and nonverbal IQ resulted in samples that may not be representative of the general population of DHH children, as considerable variability exists in terms of DHH children's cognitive and language abilities. However, for these tests the ASL team wanted to have normative samples that represent what achievements are possible when children have early and rich full access to language. In this way, the children tested will be compared to a normative group that is acquiring ASL age-appropriately, and, if delayed, can be provided with the necessary supports to develop their skills to their full potential.

Table 4. Description of normative sample for ASL adaptation

		Gender		Parents	
Age (years)	**n**	**Male**	**Female**	**Deaf**	**Non-Deaf**
3.5–4.9	25	8	15	14	11
5.0–5.9	23	10	13	15	8
6.0–6.9	23	9	14	14	9
7.0–8.9	50	26	24	36	14
9.0+	94	38	56	65	29
Totals	**215**	**125**	**90**	**144**	**71**

Recruitment for the DGS test normative sample involved alternative recruitment procedures to contact as many native signing DHH children as possible. On the one hand, the Deaf community was asked for participation of children that are native signers through an advertisement in the national magazine for the Deaf, contact to parents' associations, as well as through private contacts. On the other hand, parents of DHH children that had access to DGS from early childhood were asked for their child's participation through special schools in many different parts of Germany.

Only children with access to DGS from birth (native signers) or early childhood, i.e. before primary school, and without additional disabilities were eligible to participate in the study. Due to financial restrictions and the organizational burden of collecting data from all across Germany, the DGS team was not able to conduct additional intelligence tests but had to rely on answers in the parent and educator questionnaires. The questionnaire contained questions about additional disabilities as well as the results of the latest intelligence testing, or, if not available, the educators' assessment of the possibility of reduced intellectual capability. As a result, the DGS team was able to recruit 97 children from various regions in Germany. The variations in their linguistic background are shown in Table 5.

Table 5. Description of normative sample for DGS adaptation

Age (year;months)	n	Gender		Child DHH			
		Male	Female	Parents DHH DGS	Parents DHH no DGS	Parents hearing	CODAs DGS
4;00–4;11	12	7	5	10	0	1	1
5;00–5;11	9	4	5	3	1	2	3
6;00–6;11	13	7	6	8	1	4	0
7;00–7;11	13	10	3	9	2	2	0
8;00–8;11	12	7	5	8	0	3	1
9;00–9;11	14	8	6	11	0	3	0
10;00–10;11	16	8	8	11	0	3	2
11;00–11;11	8	5	3	5	1	2	0
Total	**97**	**56**	**41**	**65**	**5**	**20**	**7**

The quantitative study was conducted at different times in various locations across the country. The majority of participating children ($n = 65$) are native signers. These 65 children are DHH, have parents that are DHH and use DGS as their family language. Additionally, five children are DHH and have parents that are DHH but do not use DGS as their family language. Twenty children are DHH with hearing parents. The children that are not native signers have access to DGS at the latest from kindergarten and communicate in DGS. As in the normative sample of the *BSL PT*, seven native-signing hearing children, CODAs (children of Deaf adults), were included. These children have parents that are DHH and use DGS as their family language. Although in some studies, hearing native signers are used as a control group, they were included here as native signers. After sample collection, the DGS team checked whether the results of the CODAs had a significant effect on the study results compared to the results of the group of DHH children. The effect of hearing status, DHH or hearing, of the native signers on the test result was estimated using a GAM (Generalised Additive Model, Hastie & Tibshirani 1990: 169) in R (R Core Team 2019). No significant difference was found ($p = 0.99$). With 97 participants, this is the largest study of children's DGS-production so far conducted in Germany.

5.2 Terminology

Sign language assessment is a relatively new field and requires input from various disciplines, including linguistics, education, and psychology. Each of these disci-

plines brings a particular perspective which can influence the use, interpretation and understanding of terms. Specifically, we encountered some issues and confusion in determining what terms to use when discussing the grammatical features of sign languages.

Standard linguistic terms are often applied to sign languages to emphasize equal status with spoken languages and to ensure they are regarded as full-fledged natural languages (Baker, van den Bogaerde, Pfau & Schermer 2016). Often the concepts are similar, so using the same terms makes sense. For example, *semantics* can reference meaning expressed in various forms and modalities. However, there are unique differences and features of sign languages that require terms not used to describe spoken languages. For example, *embodiment* or *constructed action*, depicting the actions or feelings of characters with the hands, face and body, is an important component of sign languages to mark relationships, interaction and dialogue (Aarons & Morgan 2003; Dudis 2004).

The linguistic terms used to mark the unique aspects of sign language grammar continue to evolve and are still being identified. The use of space to establish reference points for people and objects that are not present in the conversational setting presents us with constructs that are shared between spoken and signed languages, like pronouns and agreement verbs (Valli, Lucas, & Mulroney 2005), but also with unique constructs, like spatial verbs and constructed action (Lillo-Martin 2012; Schick 2010). Linguists are analyzing some of the initial terms that encompassed numerous grammatical functions, like *classifiers*, and are separating them into more sophisticated categories to distinguish specific features, introducing terms like *spatially modifiable verbs* and *depiction* (Cormier, Smith & Sevcikova 2013; Quinto-Pozos 2007). There are linguists that emphasize the importance of developing new linguistic models that move away from spoken language traditions to make room for the modality-specific structures of sign languages (Liddell 2003). They propose using new strategies in analyzing linguistic data (Johnston & Schembri 2007).

Such linguistic debates are the source of many questions and difficult decisions during an adaptation process. Which linguistic categories can still be maintained and which should be altered based on recent developments are important considerations. For example, Schembri, Cormier & Fenlon (2018) questioned the distinction between spatial verbs and agreement verbs and proposed a new Construction Grammar analysis. Also, moving the evaluation category of role shift to constructed action affects the items in the category of manner, since it raises the question of why those items are not viewed as part of a parallel constructed action (Fischer & Kollien 2006: 457).

The solution for the ASL test (named *ASL Expressive Skills Test, ASL-EST*) was to widely maintain the terminology of the BSL test. The research team of the DGS test decided to change some terms (e.g., constructed action for role shift).

Table 6. List of terminology used in BSL PT, ASL EST, and NaKom DGS

TERMINOLOGY	
BSL PT and ASL EST	**NaKom DGS**
Narrative content	German: Erzählinhalt (translated: narrative content)
Narrative structure	German: Erzählstruktur (translated: narrative structure)
Spatial verbs	German: Abbildende Modifikationen (translated: depicting modifications)
Agreement verbs	German: Direktionale Modifikationen (translated: directional modifications)
Aspect	German: Aspektuelle Modifikationen (translated: aspectual modifications)
Manner	German: Modifikationen der Art und Weise (translated: modifications of manner)
Role shift	German: Konstruierte Aktion (translated: constructed action)

An additional layer in selecting test terminology is that differences exist in how information can be expressed in the spoken/written languages associated with the different sign languages. In our case, this refers to terminology in English and German and the fact that equivalent translations between these two languages were not always possible or feasible. For example, currently there is no commonly-used German translation of the term "constructed action", the English term is used in the DGS linguistic debate. Furthermore, the concepts associated with the linguistic discussion in one language are not always the same in the linguistic discussion in another language. For instance, in sign language linguistics, how to analyze constructions like $_{1}$GIVE$_{3}$ is debated (see Figure 2).

Cognitive-functionalistic linguists argue that the directionality involved is a fusion of signs and gestures (Schembri, Cormier & Fenlon 2018) whereas formal linguists would analyze it morpho-syntactically as verb agreement (Lillo-Martin & Meier 2011). Although "directional verbs" is a term in English associated with a formalist view, in contrast, Erlenkamp (2012) proposed to use "Richtungsverben",

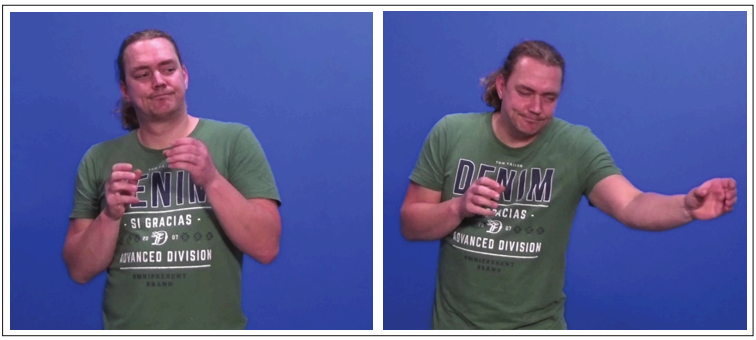

Figure 2. Pictures from the DGS model version as example for $_{1}\text{GIVE}_{3}$

which is the translation of "directional verbs" in German, as a theory-neutral description.

While we recognize that more accurate reflection of grammatical function is important, we also want to develop sign language assessment tools that are relevant and meaningful for teachers. For this reason, the ASL team chose to keep some of the more general terms within the tests (for example, *handling classifiers, role shift*), which may not be as linguistically distinctive or current, but are helpful categories and terms for teachers in responding to students' learning needs. The DGS team decided to modify the spatial verbs category, the category using *handling classifiers* as assessment criteria, to "abbildende Modifikationen" (depicting modifications of verbs).

We need to be aware of the linguistic categories that professionals like teachers or therapists already know. It is also favorable that publications about these linguistic categories are already available in written German or signed in DGS to provide professionals with extra background information, without them additionally having to struggle with input in a foreign language. A very important question is also which concepts and labels are used in the Deaf community. When considering the various aspects of linguistic terminology, the translation choices are limited and, therefore, adaptation to more recent linguistic concepts becomes questionable. This is particularly the case because linguistic concepts continue to be debated in sign language linguistic research. Although the *BSL PT* was published in 2004, which might seem to be a long time ago for a language testing tool, some of the concepts originally tested in the *BSL PT* (Herman et al. 2004) are still being discussed in sign language linguistics.

5.3 Revisions to scoring system

During the initial phases of adapting the *BSL PT* for use in ASL and DGS, the original BSL scoring system was maintained. This intuitively made sense for the purpose of comparing children's acquisition patterns across different sign languages, and all tests were assessing the same grammatical structures. However, as more data were collected and analyzed with the ASL and DGS versions, the scoring did not accurately reflect or distinguish some of the children's linguistic behaviors observed by the examiners. This created quite a dilemma, and the decision to proceed with revising the scoring system in the ASL and DGS versions of the test was not made lightly. Although these revisions do not allow for a direct comparison between the three tests, they do provide more accurate information and guidance regarding children's levels of expressive sign language skills. The two major revisions made in the ASL scoring system included combining Narrative Structure and Narrative Content and expanding the ranking levels for mastery of role shift/constructed action. In the DGS test, only the latter was changed.

Combining narrative structure and content

The ASL examiners observed significant redundancy across the two scoring categories of Narrative Structure and Narrative Content. Narrative Structure is based on high point analysis (Labov & Waletsky 1967), and whether the child's narration includes the elements of orientation (story setting), complicating actions, climax, resolution, evaluation and sequence. Narrative Content simply looks at how many of the 16 episodes are included in the narration of the story. Many of these episodes also convey the structural elements. So, for example, if the child includes the episodes that are part of complicating actions, they receive points for both content and structure when expressing the same statements. There are benefits to separating the analysis of narrative structure and content; however, the ASL researchers felt that the duplication in scoring resulted in an over-emphasis on narrative abilities in comparison to grammar. The decision to combine the narrative content and narrative structure categories was based on evidence from initial pilot testing ($n = 47$), where none of the children's narrative structure score ever exceeded their narrative content score, indicating that structures were expressed through content/episodes.

The process of combining narrative structure and content was accomplished by using the same high point analysis structure in the ASL test but adding additional points for the elements that must be expressed through several episodes. The same narrative structures/content are still shared across the three tests, which reflects overall language acquisition; however, the point values differ in the ASL version.

The distribution of narrative and grammar components is also reflected in the total maximum points assigned across the two tests. The total maximum score for the *BSL PT* is 64, with sub-totals of 34 for narrative abilities (structure – 12; content – 22) and 30 for grammar. In order to emphasize grammatical assessment and reduce redundancy in the narrative categories, the *ASL-EST* has a total maximum score of 60, with subtotals of 24 for narrative abilities and 36 for grammar.

In the German adaptation process, the original distinction between narrative structure and content was maintained, because narrative structure categories are universal. During the acquisition process, children can tell contents without structuring according to the story grammar (introduction, climax, solution). Furthermore, this distinction may help teachers to identify possible strengths or difficulties of children in narrative structure or regarding narrative content. This is valuable information for teachers to incorporate into their didactic strategies.

Expanding role shift / constructed action

The importance of role shift as a measure of narrative competence cannot be over-emphasized (Cormier, Quinto-Pozos, Sevcikova & Schembri 2012; Morgan & Woll 2007). However, it is also a challenge to identify and analyze how this skill develops in children, and to assess it consistently and reliably. To overcome these challenges, the *BSL PT* incorporated a global ranking system (0–4 points) where examiners identified the child's mastery across broad levels of "not present" (0), "developing" (2), and "fully developed" (4). Descriptors for determining the ranking are provided in the areas of facial expression and mannerisms, eye gaze, head and body movements, pausing, and directing signs to specific locations. Essentially, the *ASL-EST* uses a similar ranking system, based on the same descriptors; however, ranking is divided into three related skill areas rather than one global ranking, and the total maximum score is 6, rather than 4 points. The revision was made to increase the point value of this test item due to the significant impact role shift has on effective narrative abilities.

The purpose of ranking three separate areas, including (a) facial expression/mannerisms, (b) head/shoulder shift, and (c) eye gaze/spatial location, was to establish a more consistent and reliable measurement of this complex skill. So, for example, the categories of "not present" (0), "developing" (1) and "fully developed" (2) are applied to the child's use of facial expression and mannerisms. Many children imitate the facial expressions of the characters (ranking of 0), but that is not the same as using facial expression to show the perspectives of different characters by appropriately identifying them and shifting between them (ranking of 2). Similarly, it is helpful to distinguish between a child using shoulder shift to indicate only one character (ranking of 1), in comparison to shifting consistently and clearly between both characters (ranking of 2). The use of eye gaze or spatial

referencing can be subtle, but when mastered indicates a high level of narrative and linguistic abilities. In this way, breaking down each skill for separate rankings, resulted in more reliable and accurate scoring by the ASL examiners. Effective use of role shift is such an important predictor of sign language proficiency, that it was considered critical to give this item additional scoring guidelines and point value. Findings can help teachers provide necessary guidance and scaffolding to help students develop more effective role shifting and overall narrative abilities. The score sheet for role-shift of the *ASL-EST* (Figure 3) was also implemented in the DGS-test.

Item substitution

During the DGS pre-test, none of the deaf adults narrating the original input video or the newly produced German input video signed the item of the BSL Production Test in the manner category "(GIRL) HUNGRY/THIRSTY$_{\text{intensifier}}$" ▶, although it is possible to produce this content in DGS in a construction that fits with the description used for the BSL evaluation (see additional material).

In order to keep the balance between the number of points in each of the test categories, a substitute item in the manner category was introduced. The substitute item is "(GIRL) ANNOYED/ANGRY$_{\text{intensifier}}$" ▶ (see additional material).

During the following test adaptation process, both items were scored to see whether children might use the original item or the substitute item in their narrations. In the DGS stories, children did not sign the original item of the BSL test but did sign the newly introduced item. The item was therefore considered to be an appropriate substitution.

An item from the *BSL PT* that was produced inconsistently by the Deaf adults in the ASL pre-test was the narrative content item "boy throws spider". This may have been a result of the revised video, where this final throwing action was not as prominent as in the original video. The ASL scoring system was revised to require only the narrative content of "boy chases girl" to indicate the story resolution. In addition, some items across the three ASL story videos were not identical, but items were balanced within test categories. For example, an item in the "Aspect" category to assess "durative aspect" is indicated in one story with "(BOY) WATCH-TV" ▶ and in another story with "(GIRL) READ-MAGAZINE" ▶ (see additional material).

Recording systems

Both test adaptations changed the design of test materials from paper evaluation sheets to digital excel sheets. The ASL excel sheet was designed to be easily computer-readable (see Figure 4), whereas in the design of the DGS excel sheet,

Name: _____ Age: _____

SPIDER: Role shift

	A. ROLE SHIFT-NOT YET DEVELOPED	B. ROLE SHIFT-INCONSISTENT	C. ROLE SHIFT-FULLY DEVELOPED
A: Facial expression & mannerisms:	Some use of general facial expression	Use of facial expression to show roles, but inconsistent	Facial expression consistent with two characters
Evidence for development of role shift: Check which observed: ✓			
Boy's menacing expression, Girl's fed up expression, Boy's change of expression when eating spider			
Actual score	0	1	2
Child's score			
B: Shoulder/ Head shifting	May act out some of the story, but no use of shoulder/head shifts to show different characters	Inconsistent use of head/shoulder shifts; and/or identify characters by "name" rather than reference point/location	Clearly and consistently establishing characters and shifting roles in the story
Evidence for development of role shift: Check which observed: ✓			
Boy demanding (various items) , Girl giving (various items), Narrating the story (e.g., getting food; the chase)			
Actual score	0	1	2
Child's score			
C: Eye gaze and spatial location	Characters/spatial locations not clearly defined	Inconsistently setting up spatial locations for characters, and inconsistently referring back to locations with points/eye gaze/ direction of signs	Setting up spatial locations for both characters, and appropriately referring back to these locations with points/eye gaze/ direction of signs
Evidence for development of role shift: Check which observed: ✓			
Girl looks toward/away from boy and shakes head, Girl sees spider then looks around innocently, Girl looks towards the boy when about to eat sandwich & hands over sandwich			
Actual score	0	1	2
Child's score			
Role shift total (6)	0		

Figure 3. ASL role shift score sheet

the focus was user-orientation (see Figure 5). It includes notes regarding the evaluation criteria as in the BSL original. These scoring criteria are provided in a separate document in the ASL test.

Aspect:	Child's retell checklist	Actual score	Child's score
(E) EAT-X			
CANDY		1	
CUPCAKE			
(BOY) DRINKING		1	
Total		**2**	**0**

Manner:	Child's retell checklist	Actual score	Child's score
(GIRL) NO		1	
(BOY) DEMAND		1	
(GIRL) LOOK-AROUND		1	
(BOY) CHEW		1	
Total		**4**	**0**

Figure 4. Example of ASL score sheet

Modifications All criteria must be met.									
Aspect						**Subtotal aspect**	0		
EAT-inhibited/inceptive AS1		(GIRL) DRINK-ineeptive AS2		(BOY) WATCH-durative AS3		(BOY) DRINK-durative AS4		TAKE TAKE/DEM AND DEMAND-habitual AS5	
movement of EAT towards mouth		movement of DRINK towards mouth		slow/repeated movement of WATCH		slow, extended movement of DRINK towards mouth		repeated movement of TAKE or DEMAND	
hold		hold		possibly with hold at the end of the movement					
Comments:		Comments:		Comments:		Comments:		Comments:	

Figure 5. Example of DGS score sheet

Currently, online-only access of tests is very much en vogue. In this format, accessibility is only through an official website, and each test use needs to be paid for individually. Despite the advantages of online tests, including reduced initial costs for test users and ongoing income for test producers, both ASL and DGS research teams decided to make test material available on an USB-drive. A key reason for this decision is that reliable online access for all test users is not plausible, especially in remote areas of Canada, but also due to overloaded school networks in Germany. Feedback from the BSL test developers also indicated that training school personnel to analyze signed narratives was beneficial in promoting improved understanding of and instruction in sign language development.

6. Conclusion

We conducted our adaptations of the *BSL PT* separately, but through discussions of our experiences, and we realized that working collaboratively can enhance the test adaptation process and result. This was the motivation for writing this paper – to share our adaptation process and framework so that others may benefit from what we learned. Based on our experiences of adapting the *BSL PT* for use in ASL and DGS, we conclude that test adaptation is complex, but worthwhile. We also emphasize that it is important to follow a sound scientific and ethical process. Procedures must be put in place that prevent bias during the test adaptation process, including construct, method, and item biases. In particular with sign languages, as they represent cultural and linguistic minorities, native users of the languages must be consulted and actively involved in the test adaptation and development process.

Furthermore, test development always depends on the availability of resources (personnel, finances) in the country. In Germany, test adaptation was only possible in the context of dedicated PhD projects. This also leads to differences in the adaptation process, e.g., number of the normative sample, feasibility of conducting a pilot study.

For sign languages that do not have a large population of users, it can be particularly beneficial to adapt the established formats and developmental sequences that are present in existing tests. International comparison can facilitate evaluation and support credibility of studies with small sample sizes and different resources.

Further international studies are needed, as test results of the same tests adapted for different sign languages can provide insights into the sign language development of children. The focus of this paper was to outline our adaptation processes. Once our test results and analyses are finalized, the complete study data will be discussed in a follow-up article, and we will compare our findings across sign languages. There is a need for future research to determine in how far children in different countries, using different sign languages, acquire these languages in similar and different ways.

Funding

International collaboration was facilitated through the Visiting Scholar program of DAAD (*Deutscher Akademischer Austausch Dienst*) – German Academic Exchange Service.

Acknowledgements

We thank the participating children for their willingness to participate in the research as well as for their excellent sign language narrations. We are also very thankful to the parents and teachers for their consent and great support. A test adaptation is a lengthy process that can only succeed with the help of a dedicated research team.

The video clips and photos in DGS are part of the model version signed by Tino Sell. The video clips in ASL were signed by research assistant, Kyra Zimmer.

References

Aarons, Debra & Ruth Morgan. 2003. Classifier predicates and the creation of multiple perspectives in South African Sign Language. *Sign Language Studies* 3(2). 125–156. https://doi.org/10.1353/sls.2003.0001

Anderson, Diane. 2006. Lexical development of deaf children acquiring signed languages. In Brenda Schick (ed.), *Advances in the sign language development of deaf children*, 135–160. New York, NY: Oxford University Press.

Baker, Anne, Beppie van den Bogaerde, Roland Pfau & Trude Schermer. 2016. *The linguistics of sign languages: An introduction*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.199

Becker, Claudia. 2009. Narrative competences of Deaf children in German Sign Language. *Sign Language & Linguistics* 12(2). 113–160. https://doi.org/10.1075/sll.12.2.02bec

Becker, Claudia, Martje Hansen & Patricia Barbeito Rey-Geissler. 2018. Narrative Kompetenzen hörgeschädigter Kinder – Die Interaktion von Gebärdenspracherwerb und Theory of Mind. *Das Zeichen* 108. 90–105.

Boudreault, Patrick, Kyra Zimmer & Charlotte Enns. 2015. Creating videos to assess children's signed language narrative skills. *Proceedings of the 22nd International Congress on the Education of the Deaf (ICED 2015)*. Athens, Greece, 6–9 July.

Brown, Linda, Rita J. Sherbenou & Susan K. Johnsen. 2010. *Test of nonverbal intelligence, 4th edition*. New York, NY: Pearson.

Chen Pichler, Deborah. 2012. Acquisition. In Roland Pfau, Markus Steinbach & Bencie Woll (eds.), *Sign language: An international handbook*, 647–686. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110261325.647

Cokely, Dennis & Charlotte Baker. 1991. *American Sign Language: A teacher's resource text on grammar and culture*. Washington, DC: Gallaudet University Press.

Cormier, Kearsy, David Quinto-Pozos, Zed Sevcikova & Adam Schembri. 2012. Lexicalisation and delexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language & Communication* 32. 329–348. https://doi.org/10.1016/j.langcom.2012.09.004

Cormier, Kearsy, Sandra Smith & Zed Sevcikova. 2013. Predicate structures, gesture, and simultaneity in the representation of action in British Sign Language: Evidence from deaf children and adults. *Journal of Deaf Studies and Deaf Education* 18(3). 370–390. https://doi.org/10.1093/deafed/ent020

Council of Europe. 2001. *Common European Framework of References for Languages: Learning, teaching, assessment*. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680459f97. Access date 2020-03-02.

Cravens, Elisabeth. 2013. *Evaluating the utility of the Test of Narrative Language for use with deaf children via American Sign Language*. Austin, TX: University of Texas Master's thesis.

Dudis, Paul. 2004. Body partitioning and real-space blends. *Cognitive Linguistics* 15(2). 223–238. https://doi.org/10.1515/cogl.2004.009

Dunn, Lloyd M. & Douglas M. Dunn. 1997. *Peabody picture vocabulary test (3rd ed.)*. Circle Pines, MN: American Guidance Services.

ELAN (Version 5.5) [Computer software]. (2019). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Emmorey, Karen & Judy Reilly. 1998. The development of quotation and reported action: Conveying perspective in ASL. In Eve V. Clark (ed.), *Proceedings of the Twenty-ninth Annual Stanford Child Language Research Forum*, 81–90. Stanford, CA: CSLI Publications.

Enns, Charlotte, Kyra Zimmer, Patrick Boudreault, Sarah Rabu & Cheryle Broszeit. 2013. *American Sign Language Receptive Skills Test*. Winnipeg, MB: Northern Signs Research. www.northernsignsresearch.com

Enns, Charlotte, Tobias Haug, Rosalind Herman, Robert Hoffmeister, Wolfgang Mann & Lynn McQuarrie. 2016. Exploring signed language assessment tools around the world. In Marc Marschark, Venetta Lampropoulou & Emmanouil K. Skordilis (eds.), *Diversity in deaf education*, 171–218. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190493073.003.0007

Enns, Charlotte, Kyra Zimmer, Cheryle Broszeit & Sarah Rabu. 2019. *American Sign Language Expressive Skills Test*. Winnipeg, MB: Northern Signs Research.

Erlenkamp, Sonja. 2012. Syntax: Aus Gebärden Sätze bilden. In Hanna Eichmann, Martje Hansen & Jens Heßmann (eds.), *Handbuch Deutsche Gebärdensprache: sprachwissenschaftliche und anwendungsbezogene Perspektiven*, 165–198. Seedorf: Signum-Verlag.

Fischer, Renate & Simon Kollien. 2006. Constructed action in DGS: Roses Aktions = Fragmente (Teil II). *Das Zeichen* 74. 448–463.

Gagarina, Natalia, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Ute Bohnacker & Joel Walters. 2012. MAIN – Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics* 56. 1–155.

Hall, Wyatte. 2017. What you don't know can hurt you: The risk of language deprivation by impairing sign language development in Deaf children. *Maternal and Child Health Journal* 21(5). 961–965. https://doi.org/10.1007/s10995-017-2287-y

Hänel, Barbara. 2005. *Der Erwerb der Deutschen Gebärdensprache als Erstsprache: die frühkindliche Sprachentwicklung von Subjekt- und Objektverbkongruenz in DGS*. Tübingen: Narr.

Happ, Daniela & Marc-Oliver Vorkörper. 2014. *Deutsche Gebärdensprache: Ein Lehr- und Arbeitsbuch*. Frankfurt: Fachhochschulverlag.

Harris, Raychelle, Heidi Holmes & Donna Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies* 9(2). 104–131. https://doi.org/10.1353/sls.0.0011

Hastie, Trevor & Robert Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Haug, Tobias & Wolfgang Mann. 2008. Adapting tests of sign language assessment for other sign languages – A review of linguistic, cultural, and psychometric problems. *Journal of Deaf Studies and Deaf Education* 13(1). 138–147. https://doi.org/10.1093/deafed/enm027

Haug, Tobias. 2011. Approaching sign language test construction: Adaptation of the German Sign Language receptive skills test. *Journal of Deaf Studies and Deaf Education* 16(3). 343–361. https://doi.org/10.1093/deafed/enq062

Henner, Jon, Catherine Caldwell-Harris, Rama Novodgrodsky & Robert Hoffmeister. 2016. American Sign Language syntax and analogical reasoning skills are influenced by early acquisition and age of entry to signing schools for the deaf. *Frontiers in Psychology* 26. https://doi.org/10.3389/fpsyg.2016.01982

Herman, Rosalind. 1998. The need for an assessment of deaf children's signing skills. *Deafness and Education: Journal of the British Association of the Teachers of the Deaf* 22(3). 3–8.

Herman, Rosalind. 2015. Language assessment of Deaf learners. In Harry Knoors & Marc Marschark (eds.), *Educating Deaf learners: Creating a global evidence base*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190215194.003.0009

Herman, Rosalind, Sallie Holmes & Bencie Woll. 1999. *Assessing BSL development: Receptive skills test*. Coleford, UK: Forest Bookshop.

Herman, Rosalind, Nicola Grove, Sallie Holmes, Gary Morgan, Hillary Sutherland & Bencie Woll. 2004. *Assessing BSL development: Production test (narrative skills)*. London, UK: City University.

Hoffmeister, Robert, Catherine Caldwell-Harris, Jonathan Henner, Rachel Benedict, Sarah Fish, Patrick Rosenburg, Frances Conlin-Luippold & Rama Novogrodsky. 2014. *The American Sign Language Assessment Instrument (ASLAI): Progress report and preliminary findings*. Working paper. Boston, MA: Center for the Study of Communication and the Deaf.

Iliescu, Dragos. 2017. *Adapting tests in linguistic and cultural situations*. New York, NY: Cambridge University Press. https://doi.org/10.1017/9781316273203

International Test Commission. 2017. *The ITC guidelines for translating and adapting tests (Second edition)*. Retrieved from www.InTestCom.org/page/16. Access Date 2020-06-02.

Johnston, Trevor. 2004. The assessment and achievement of proficiency in a native sign language within a sign bilingual program: The pilot Auslan Receptive Skills Test. *Deafness and Education International* 6(2). 57–81. https://doi.org/10.1179/146431504790560582

Johnston, Trevor & Adam Schembri. 2007. *Australian Sign Language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511607479

Jones, Neil. 2012. Reliability and dependability. In Glenn Fulcher & Fred Davidson (eds.), *The Routledge handbook of language testing*, 350–362. London, UK: Routledge.

Klima, Edward & Ursula Bellugi. 1979. *The signs of language*. Cambridge, MA: Harvard University Press.

Kolbe, Vera. 2019. Beobachtungsverfahren für Deutsche Gebärdensprache. *Hörgeschädigtenpädagogik* 3/2019. 169–174. https://doi.org/10.18452/22201

Levshina, Natalia. 2015. *How to do linguistics with R: data exploration and statistical analysis*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.195

Liddell, Scott. 2003. *Grammar, gesture, and meaning in American Sign Language*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511615054

Labov, William & Joshua Waletsky. 1967. Oral versions of personal experiences. In June Helm (ed.), *Essays on the verbal and visual arts*, 12–44. Seattle, WA: University of Washington Press.

Lillo-Martin, Diane & Richard Meier. 2011. On the linguistic status of 'agreement' in sign languages. *Theoretical Linguistics* 37. 95–142. https://doi.org/10.1515/thli.2011.009

Lillo-Martin, Diane. 2012. Utterance reports and constructed action in sign and spoken languages. In Roland Pfau, Markus Steinbach & Bencie Woll (eds.), *Sign language: An international handbook*, 365–387. Berlin: De Gruyter Mouton.

Meier, Richard P. 1991. Language acquisition by Deaf children. *American Scientist* 79(1). 60–70.

Meier, Richard P. 2002. Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In Richard P. Meier, Kearsy Cormier & David Quinto-Pozos (eds.), *Modality and structure in signed and spoken language*, 1–25. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486777.001

Meristo, Marek, Kerstin Falkman, Erland Hjelmquist, Mariantonia Tedoldi, Luca Surian & Michael Siegal. 2007. Language and theory of mind reasoning: Evidence from deaf children in bilingual and oralist environments. *Developmental Psychology* 43. 1156–1169. https://doi.org/10.1037/0012-1649.43.5.1156

Mitchell, Ross & Michael Karchmer. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies* 4. 138–163. https://doi.org/10.1353/sls.2004.0005

Neidle, Carol, Judy Kegl, Dawn MacLaughlin, Benjamin Bahan & Robert Lee. 2000. *The syntax of American Sign Language: Functional categories and hierarchical structure*. Cambridge, MA: MIT Press.

Morgan, Gary & Bencie Woll. 2007. Understanding sign language classifiers through a polycomponential approach. *Lingua* 117. 1159–1168. https://doi.org/10.1016/j.lingua.2006.01.006

Novogrodsky, Rama & Natalia Meir. 2020. Age, frequency, and iconicity in early sign language acquisition: Evidence from the Israeli Sign Language MacArthur-Bates Communicative Developmental Inventory. *Applied Psycholinguistics* 41(4). 817–845. https://doi.org/10.1017/S0142716420000247

Oakland, Thomas & Hollie Lane. 2004. Language, reading and readability formulas: Implications for developing and adapting tests. *International Journal of Testing* 4(3). 239–252. https://doi.org/10.1207/s15327574ijt0403_3

O'Sullivan, Barry & Cyril Weir. 2011. Test development and validation. In O'Sullivan, Barry (ed.), *Language testing: theories and practices* (1st publ. ed.). Basingstoke: Palgrave Macmillan.

Papaspyrou, Chrissostomos, Alexander von Meyenn, Michaela Matthaei & Bettina Herrmann. 2008. *Grammatik der Deutschen Gebärdensprache aus der Sicht gehörloser Fachleute*. Seedorf: Signum Verlag.

Quinto-Pozos, David. 2007. Can constructed action be considered obligatory? *Lingua* 117. 1285–1314. https://doi.org/10.1016/j.lingua.2005.12.003

R Core Team. 2019. R: *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Reilly, Judy. 2005. How faces come to serve grammar: The development of nonmanual morphology in American Sign Language. In Brenda Schick, Marc Marschark & Patricia Elisabeth Spencer (eds.), *Advances in the sign language development of deaf children*, 262–290. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195180947.003.0011

Reckwitz, Andreas. 2011. Die Kontingenzperspektive der ›Kultur‹. Kulturbegriffe, Kulturtheorien und das kulturwissenschaftliche Forschungsprogramm. In Friedrich Jaeger & Burkhard Liebsch (eds.), *Handbuch der Kulturwissenschaften: Grundlagen und Schlüsselbegriffe*, 1–20. Stuttgart: J.B. Metzler. https://doi.org/10.1007/978-3-476-05012-0_1

Schembri Adam, Kearsy Cormier & Jordan Fenlon. 2018. Indicating verbs as typologically unique constructions: Reconsidering verb 'agreement' in sign languages. *Glossa: a Journal of General Linguistics* 3(89). 1–40. https://doi.org/10.5334/gjgl.468

Schembri, Adam, Gillian Wigglesworth, Trevor Johnston, Greg Leigh, Robert Adam & Roz Barker. 2002. Issues in development of the Test Battery for Australian Sign Language Morphology and Syntax. *Journal of Deaf Studies and Deaf Education* 7(1). 18–40. https://doi.org/10.1093/deafed/7.1.18

Schick, Brenda. 2002. The expression of grammatical relations in deaf toddlers learning ASL. In Gary Morgan & Bencie Woll (eds.), *Directions in sign language acquisition*, 143–158. Amsterdam: John Benjamins. https://doi.org/10.1075/tilar.2.09sch

Schick, Brenda. 2010. The development of American Sign Language and Manually Coded English systems. In Marc Marschark & Patricia Elisabeth Spencer (eds.), *The Oxford handbook of deaf studies, language, and education (Volume 1, second edition)*, 229–240. New York, NY: Oxford University Press.

Schwager, Waldemar. 2012. Morphologie: Bildung und Modifikation von Gebärden. In Hanna Eichmann, Martje Hansen & Jens Heßmann (eds.) *Handbuch Deutsche Gebärdensprache: sprachwissenschaftliche und anwendungsbezogene Perspektiven*, 61–110. Seedorf: Signum-Verlag

Sign Language Linguistic Society. 2016. *Ethics statement for sign language research*. Retrieved from https://slls.eu/slls-ethics-statement/. Access date 2019-09-18.

Simms, Laurene, Sharon Baker & Diane M. Clark. 2013. The Standardized Visual Communication and Sign Language Checklist for signing children. *Sign Language Studies* 14(1). 101–124. https://doi.org/10.1353/sls.2013.0029

Singleton, Jenny L. & Gary Morgan. 2006. Natural signed language acquisition within the social context of the classroom. In Brenda Schick, Marc Marschark & Patricia Elisabeth Spencer (eds.), *Advances in the sign language development of deaf children*, 344–376. New York, NY: Oxford University Press.

Singleton, Jenny L. & Elissa L. Newport. 2004. When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology* 49. 370–407. https://doi.org/10.1016/j.cogpsych.2004.05.001

Singleton, Jenny L. & Samuel Supalla. 2011. Assessing children's proficiency in natural signed languages. In Marc Marschark & Patricia E. Spencer (eds.), *Oxford handbook of deaf studies, language, and education (Vol. 1, 2nd edition)*, 289–302. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199750986.013.0022

Valli, Clayton, Ceil Lucas & Kristin Mulroney. 2005. *Linguistics of American Sign Language: An introduction (Fourth edition)*. Washington, DC: Gallaudet University Press.

Valmaseda, Marian, Mar Pérez, Rosalind Herman, Nuria Ramírez & Ignacio Montero. September 2013. *Evaluación de la competencia gramatical en LSE: Proceso de adaptación del BSL Receptive Skill Test (test de habilidades receptivas)*. Paper presented at the Congreso CNLSE sobre la Investigación de la Lengua de Signos. Madrid, Spain.

Address for correspondence

Charlotte Enns
University of Manitoba
252 St. Paul's College
Winnipeg, Manitoba R3T 2N2
Canada
charlotte.enns@umanitoba.ca

Co-author information

Vera Kolbe
Department of Sign Language Pedagogy and Audio Pedagogy
Humboldt-Universität zu Berlin
vera.kolbe@hu-berlin.de

Claudia Becker
Department of Sign Language Pedagogy and Audio Pedagogy
Humboldt-Universität zu Berlin
claudia.becker@hu-berlin.de

Publication history

Date received: 20 August 2020
Date accepted: 24 May 2021
Published online: 19 July 2021