CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

The sentence repetition task as a measure of sign language proficiency

Krister Schönström[1]* and Peter C. Hauser[2]

[1]Department of Linguistics, Stockholm University, Stockholm, Sweden and [2]Center on Culture and Language, National Technical Institute of the Deaf, Rochester Institute of Technology, Rochester, NY, USA
*Corresponding author. Email: schonstrom@ling.su.se

Abstract
Sign language research is important for our understanding of languages in general and for the impact it has on policy and on the lives of deaf people. There is a need for a sign language proficiency measure, to use as a grouping or continuous variable, both in psycholinguistics and in other sign language research. This article describes the development of a Swedish Sign Language Sentence Repetition Test (STS-SRT) and the evidence that supports the validity of the test's interpretation and use. The STS-SRT was administered to 44 deaf adults and children, and was shown to have excellent internal reliability (Cronbach's alpha of 0.915) and inter-rater reliability (Intraclass Correlation Coefficient [ICC] = 0.900, $p < .001$). A linear mixed model analysis revealed that adults scored 20.2% higher than children, and delayed sign language acquisition were associated with lower scores. As the sign span of sentences increased, participants relied on their implicit linguistic knowledge to scaffold their sentence repetitions beyond rote memory. The results provide reliability and validity evidence to support the use of STS-SRT in research as a measure of STS proficiency.

Sign language research has significantly informed basic and applied sciences, yielding practical implications for both. For example, research on Swedish Sign Language, *Svenskt teckenspråk* (henceforth, STS), contributed to the recognition of STS as a language by the Swedish parliament in 1981 (Proposition 1980/1981:100) and the consequent implementation of sign bilingualism in the national educational curriculum. Swedish Sign Language has been a part of deaf bilingual education for nearly 40 years, both as the language of instruction for deaf pupils in classrooms, and as a school subject with its own curriculum in Swedish deaf schools (Mahshie, 1995; Svartholm, 2010). The 2009 Swedish Language Act proclaimed STS as equal to other minority languages used in Sweden and emphasized society's responsibility toward STS (SFS, 2009:600). This study describes the

CrossMark

development of the first Sentence Repetition Task (SRT)-based STS proficiency that provides a score that can be used in future studies as a continuous variable.

Psychometrically sound sign language proficiency assessments have been developed for different signed languages for first language learners (see Enns et al., 2016; Paludnevičiene et al., 2012 for a review) and second language learners (see Landa & Clark, 2019; Schönström et al., in press for a review). Hauser et al. (2008) developed ASL-SRT, for American Sign Language (ASL), based on an English oral repetition test (Hammill et al., 1994) that required participants to listen to English sentences of increasing length and morphosyntactic complexity and immediately repeat them correctly with 100% accuracy. They chose to use the SRT approach because it can be administered both to children and adults, and because past studies have shown it to be a good measure of language proficiency. Some have described the SRT as a global proficiency measure because it involves sentence processing, reconstruction, and reproduction (Haug et al., 2020; Jessop et al., 2007), while others have described it as a test of grammatical processing (Hammill et al., 1994; Spada et al., 2015) and have used it to study children's syntactic development (e.g., Kidd et al., 2007).

Spoken language SRTs are sensitive to the developmental proficiency of first language learners (e.g., Devescovi & Caselli, 2007; Klem et al., 2015) and second language learners (e.g., Erlam, 2006; Gaillard & Tremblay, 2016; Spada et al., 2015). Implicit long-term linguistic knowledge – that is, language proficiency – enhances SRT performance. When native speakers were asked to repeat ungrammatical sentences, they unconsciously applied their linguistic knowledge and corrected the grammar 91% of the time when they repeated the sentences (Erlam, 2006). This happens because individuals' language knowledge helps them to scaffold sentences in working memory so they can hold more information than just rote memory allows. Relying on rote memory alone, individuals have difficulty with sentences of more than four words (see Haug et al., 2020 and Polišenská et al., 2015 for discussion). Correct sentence repetitions have been found to positively correlate with 2–4-year-old children's mean length of utterances in free speech and verbal memory span (Devescovi & Caselli, 2007). Further, the accuracy of sentence repetitions of 4–5-year-old children depended more on their familiarity with morphosyntax and lexical phonology than on their familiarity with semantics or prosody (Polišenská et al., 2015).

Studies with both language and memory measures have shown that SRTs measure an underlying unitary language construct rather than a measure of working memory because they draw upon a wide range of language processing skills (e.g., Gaillard & Tremblay, 2016; Klem et al., 2015; Tomblin & Zhang, 2006). Sentence Repetition Test have been claimed to be the best single test for identifying children with Specific Language Impairment (SLI) due to their high specificity and sensitivity (Archibald & Joanisse, 2009; Conti-Ramsden et al., 2001; Fleckstein et al., 2016; Stokes et al., 2006). Individuals with SLI perform poorly on SRTs because the test heavily recruits linguistic processing abilities (Leclercq et al., 2014) although some have argued that SRTs test memory more than language proficiency (Borwnell, 1988; Henner et al., 2018). It has been claimed that individuals with developing language skills or SLI have less accurate sentence repetitions because they do not have available linguistic knowledge to scaffold sentences in episodic

memory (Alptekin & Erçetin, 2010; Coughlin & Tremblay, 2013; Van den Noort et al., 2006).

Hauser et al. (2008) administered the ASL-SRT to deaf and hearing children and adults and found that native signers more accurately repeated the ASL sentences than non-native signers. Supalla et al. (2014) analyzed the ASL-SRT error patterns of native signers and found that fluent signers made more semantic types of errors, suggesting top-down scaffolding mechanisms are used in working memory to succeed in the task, whereas less fluent signers made errors that were motoric imitations of signs (similar phonology) that were ungrammatical or lacked meaning. In a case study of a deaf adult native ASL signer with SLI, the participant performed poorly on the ASL-SRT compared to peers, including non-native deaf signers (Quinto-Pozos et al., 2017). The results cited above support the claim that the SRT measures sign language proficiency as well as it does spoken language proficiency.

The ASL-SRT has been adapted to several languages, including German Sign Language (DGS, Kubus et al., 2015), British Sign Language (BSL, Cormier et al., 2012), and Swiss German Sign Language (DSGS, Haug et al., 2020), among others. Other sign language tests are also sentence repetition task-based, for example, those assessing Italian Sign language (LIS, Rinaldi et al., 2018) and French Sign Language (LSF, Bogliotti et al., 2020), but those follow a different methodological framework, and development, administration, and coding procedures. This article describes how the STS-SRT was developed and outlines its psychometric properties.

Method

STS-SRT development

Translated, corpus, and novel sentences
The STS-SRT began with 60 sentences from 3 different sources; 20 were translated from the ASL-SRT (Hauser et al., 2008), 20 were from the Swedish Sign Language Corpus (STSC, Mesch & Wallin, 2012), and 20 novel sentences were developed for this project. For the first source, the original 20 ASL sentences were translated into STS by a deaf native STS signer proficient in ASL along with a trained linguist proficient in both languages. There were two reasons for translating the ASL sentences. Instead of making all sentences novel, we decided to try the sentences that have been developed and proofed for ASL-SRT, in order to keep open the possibility of future cross-linguistic comparisons. For the second source, the team extracted 20 authentic sentences from the STSC (Mesch & Wallin, 2012). The STSC consists of naturally occurring, authentic STS productions between deaf people. The motivation behind using STSC as a resource was to enhance the authenticity of the sentences. Only sentences from native STS users were selected. Efforts were made to select authentic sentences that could work context-free. Some of the sentences relate to typical topics within the deaf community; that is, basic cultural knowledge associated with being deaf. For example, one sentence contains elements from the history of oral deaf education, and another sentence describes the signing skills of teachers of the deaf. For the third source, the team also created 20 novel items. We varied items by length (short, 3–4 signs; intermediate, 5–6 signs; and long, 7–10 signs), lexical classification (e.g., lexical, fingerspelled, and classifier signs), and morphological (e.g., numeral

Sentence 6 (Translated)

SUNDAY   NEWSPAPER   ALWAYS-mod   THICK

Sentence 9 (Corpus)

SOME   TEACHER   SIGN-mod   WELL

Sentence 17 (Novel)

IF-fs   INDEX   WIN.   LOTTERY   FOR-SURE-mod   INDEX   BUY   NEW-fs   CAR
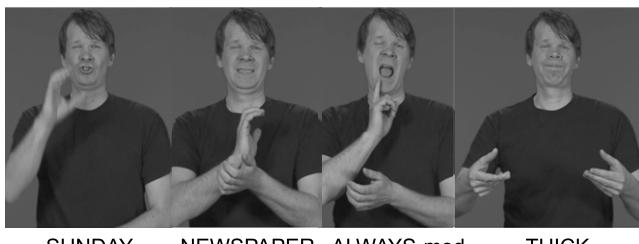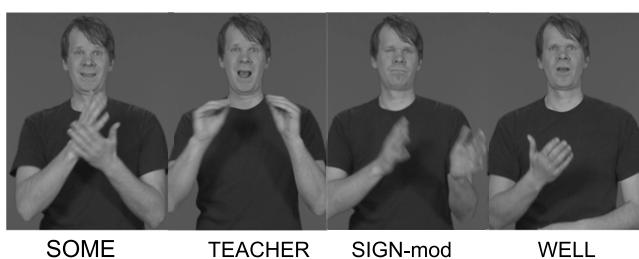
Figure 1. Three examples of STS-SRT sentences: Translated, Corpus, and Novel.

incorporation, modified signs), and syntactic (e.g., determiners, main and subordinate clauses) complexity.

A deaf native signer was filmed signing the sentences in a natural manner and tempo against a dark studio background. Care was taken to reproduce the signs precisely as produced by the authentic signers in the original ASL or STSC sentences, that is, using the same variants of signs, the same syntactic order, the same speed, etc. The 60 sentences were later presented to 3 deaf native signers for review. A criterion based on 2/3 exclusion was applied; that is, if two of the three reviewers rejected a sentence as inauthentic, inexplicit, non-fluent, or even ungrammatical, the sentence was removed. This review process left 40 sentences, of which 16 were translated sentences, 11 sentences were corpus sentences, and 13 were novel sentences (Figure 1).

Test administration procedures
A QuickTime video file was created with the STS-SRT instructions, practice sentences, and test sentences. Every sentence video clip was followed by a gray background

| | | | |
| ITEM | RESPONSE TIME | ITEM | RESPONSE TIME |

Figure 2. Format of STS-SRT sentence delivery.

for 7–9 s, depending on the sentence length; see Figure 2. There were 10–12 frames at the beginning of each sentence before the signing start time to give participants enough time to prepare themselves after pushing the space button to play the clip. Five frames were also added at the signing end time to eliminate any negative effect on memory and to give participants time to start imitating the sentences as fast as possible. The STS-SRT was presented on an Apple MacBook Pro 15 with a 15-inch screen and a built-in camera to record repetitions, in a quiet room without visual distractions. The laptop was placed on a table with a deaf research assistant sitting beside the participant.

The test begins with instructions in which the deaf signing research assistant describes the test and the procedure to be followed, that is, that the participant is going to see some sentences and will be told to repeat the sentences. Participants are instructed to repeat the sentences as seen, and are explicitly asked to use the sign produced by the signer rather than a synonym. The instructions conclude by announcing that there will be three practice sentences before the actual test begins. During the practice, the research assistant checks understanding and corrects any mistakes. During testing, the research assistant can use the space bar to pause the sentences as necessary, for example, to ensure that the participant answers within the allotted response time (7–9 s), that is, before the start of the next item. After completion, the recording is saved for later scoring.

Scoring procedures

The STS-SRT scoring procedures follow the same principles and basic instructions as the ASL-SRT (Hauser et al., 2008), that is, each correct response is scored 1 point, and each incorrect response is scored 0 points. A sentence that is not produced with the same phonology, morphology, or syntax as the stimulus sentence is considered an error and is marked as 0. Raters use a scoring protocol and mark signs in the sentences that were repeated incorrectly or omitted. The STS-SRT scoring instructions (Figure 3) include examples of errors base on phonological (e.g., different handshape used), morphological (e.g., omission of reduplication forms), and syntactic (e.g., different word order) differences. Omissions, replacements, or changes of lexical items were counted as errors. Metathesis and altered direction of signing were considered acceptable as long as they did not affect meaning.

The scoring protocol focused on manual features only, that is the signs, and not on nonmanual features, that is, use of eye gaze, eyebrows, etc., for grammatical and

11. PRO1 PERF KÖPA NY@b LÄGENHET (INDEX FINISH BUY NEW-fs APARTMENT)		
	Aktören (Actor)	Acceptabla varianter (Acceptable variants)	Oacceptabla varianter (Unacceptable variants)
	LÄGENHET, böjd flathand, framåtriktat (APARTMENT, bend D-hand, forward direction)	LÄGENHET, flathand, annan orientering (APARTMENT, D-hand, different direction)	
	NY@b (NEW-fs)		NY (NEW, no fingerspelling)

12. POSS PAPPA GILLA LÄSA BOK-pl (MY FATHER LOVE READ BOOK-pl)		
	Aktören (Actor)	Acceptabla varianter (Acceptable variants)	Oacceptabla varianter (Unacceptable variants)
	LÄSA, två händer med kontakt (READ, two-handed with contact)	LÄSA, två händer utan kontakt (READ, two-handed without contact)	LÄSA en hand (READ, one-handed)
	GILLA med G-hand (LOVE, G-hand)		GILLA med B-hand (LOVE, B-hand)

Figure 3. Two example sentences from the Scoring Instruction Manual, Original with English Translation in parenthesis.

prosodic as well as pragmatic purposes, with a few exceptions: the use of another mouthing with a completely different meaning in otherwise ambiguous sign pairs, for example, the mouthing /fa/ as in FABRIK (*factory*) where appropriate mouthing for the target item FUNGERA (*working*) would have been /fu/ even though the manual sign for FUNGERA and FABRIK is the same, but with different mouthings depending on the meaning. Additionally, nonmanual markers such as headshaking were required for the accurate production of a negative utterance.

The scoring instructions describe the point system, correct versus error responses (e.g., omission, replacement, change in word order), and examples of acceptable/ unacceptable response variants for each item. It is important that the raters are fluent STS signers and have basic knowledge of sign language linguistics in order to accurately rate the responses. We created test administration instructions aimed at test instructors, as well as scoring instructions and sheets for the raters to use. It is recommended that new raters practice with experienced raters to ensure confidence about the procedure.

STS-SRT pilot testing
The purpose of pilot testing was to review the sentences and to practice some testing in a real-life situation to examine the test's functionality. In the pilot testing, we also

observed some recurring phonological variants in signs to use in the scoring instructions, that is, a list of acceptable phonological variants on signs. A purposeful sample of 10 participants was chosen according to different background variables: four participants were deaf adults of deaf signing parents, five participants were deaf adults of hearing parents, and one was a hearing skilled signer of hearing parents. The parental hearing status of the adult participants is important because, in the absence of proficiency tests, researchers have identified deaf individuals as native signers if they were raised by deaf signing parents; they are classified as non-native signers if they have hearing parents (see Humphries et al., 2014 for discussion on the linguistic experience of deaf children of hearing parents). The pilot sample included both native and non-native signers.

Data from the pilot study helped to adjust the acceptable phonological variants according to STS linguistic rules and usage. For example, the movement of the sign LÄSA (*read*) in sentence number 12 in Figure 3 was shown to vary widely between signers, that is, alternating between a variant with movement with contact (which was the target form as produced by the signer) and a variant without. The variant was considered acceptable if more than 40% of signers in the pilot test group exhibited it.

STS-SRT final version

The number of sentences was reduced from 40 to 31 based on the pilot testing results. In the final version of the STS-SRT, 11 were translated sentences, 8 were corpus sentences, and 12 were novel sentences. Table 1 provides an overview of the 31 sentences and their linguistic properties. It was important that the remaining sentences included a wide variety of different linguistic properties. For lexicon and morphology, different types of signs such as lexical signs (including fingerspelled signs) and classifier signs were included. We also included numeral incorporations, as well as signs modified to indicate plurality, aspect, and spatial reference. For syntax, we aimed to have a variety of main and subordinate clauses, as well as different kinds of phrasal structures (e.g., determiners in noun phrases, verb phrases with and without auxiliary verbs). Subordinate clause types included relative, conditional, object, consecutive, as well as causal. In addition, clauses were presented in various word orders depending on information structure, and included such components as topic clauses, adverbial fronting, etc. Some sentences were constructed with simple syntax and simple lexical and morphological content, whereas some sentences have a complex combination of both, or a combination of both simple and complex constructions.

Sentences which no one repeated correctly in the pilot version were omitted, as were some simple sentences that everyone repeated correctly. However, anticipating a possible application to children, we decided to keep a few more of the simple sentences in order not to make the test too difficult. In total, one simple sentence, two intermediate sentences, and six complex sentences were omitted. The sentences in the final version were ordered beginning with the items most of the pilot participants repeated correctly and ending with the items that the fewest repeated correctly. Correct sentence repetitions are awarded one point, allowing for a maximum STS-SRT score of 31. The test administration time is 15 min per participant, and rating time varies between 15 and 20 min.

Table 1. Linguistic properties of the 31 STS-SRT sentences

			Description of linguistic properties	
Item	# Signs	Source	Lexicon and morphology	Syntax and discourse
1	3	Translated		Use of referential space, negation,
2	5	Novel	Modified sign, fingerspelled sign	Determiner in NP
3	5	Novel	Fingerspelled sign	Topic clause,
4	6	Novel		Auxiliary verb
5	4	Translated		Relative clause, use of referential space
6	4	Translated	Classifier sign	Topic clause, determiner in NP
7	4	Novel	Modified sign	Relative clause, use of referential space
8	4	Novel		Use of referential space
9	4	Corpus	Modified sign, compound sign	
10	4	Corpus		Adverbial-fronted sentence
11	5	Translated	Fingerspelled sign	Auxiliary verb, determiner in NP,
12	5	Translated	Plural incorporation	Object clause,
13	6	Corpus		Determiners in NP, auxiliary verb, referential space
14	5	Novel	Compound noun, Number incorporation	Topic clause, determiners in NP,
15	6	Translated	Classifier sign	Noun classifier, determiner in NP
16	5	Corpus	Indicating sign, palm-up gesture	Causal clause, auxiliary verb,
17	9	Novel	Fingerspelled sign, modified sign,	Conditional clause, determiner in NP
18	6	Corpus	Indicating signs	Referential space,
19	4	Novel		Topic clause, negation, pronominal copy
20	5	Novel		Conditional clause
21	7	Corpus	Fingerspelled sign, indicating sign	Topic clause, determiner in NP
22	10	Translated	Modified sign	Relative clause, determiner in NP
23	8	Novel	Compound sign	Object clause, auxiliary verb
24	8	Novel	Classifier sign, fingerspelled sign	Relative clause, determiner in NP
25	5	Novel	Indicating verb, number incorporation	
26	8	Translated	Number incorporation, modified sign	Auxiliary verbs, serial verbs, determiner in NP

(*Continued*)

Table 1. (*Continued*)

Item	# Signs	Source	Lexicon and morphology	Syntax and discourse
			Description of linguistic properties	
27	7	Translated	Classifier signs	Consecutive clause, determiners in NP
28	9	Translated	Classifier signs, fingerspelled sign, modified sign	Consecutive clause, determiner in NP
29	9	Corpus	Compound signs, fingerspelled signs	Adverbial fronting, determiner in NP
30	8	Corpus	Indicating sign	Auxiliary verb
31	7	Translated	Modified sign	Consecutive clause, determiner in NP

Table 2. Descriptive statistics of test participants

Group	*n*	Age range	Mean age	*SD*	*n* Deaf parents
Adults	30	20–49	32.23	7.94	11
Children	14	10–16	12.79	2.36	6

STS-SRT psychometric testing

Sample

There were 44 participants in the psychometric testing sample, including 14 deaf children (8 females and 6 males) and 30 deaf adults (17 females and 13 males); see Tables 2 and 3 for descriptive statistics and overview of the participants. Adult participants were asked to fill out a background questionnaire in which they answered questions about their language, identity, education, hearing status, parents' hearing status and language background, and age they were first exposed to STS (Age of Acquisition [AoA]). Background questionnaires of child participants were filled by their parents. The study procedures were approved by the Swedish Ethical Board.

Results

Reliability evidence

A deaf fluent signer, who is a trained linguist, rated all of the 44 participants' STS-SRT repetitions and a second rater, also a deaf fluent signer and trained linguist, independently rated the repetitions of a subsample of 29 participants to determine inter-rater reliability. Inter-rater reliability was computed between the raters using the Intraclass Correlation Coefficient (ICC) with a two-way mixed-effects model for single rater protocols (Koo & Li, 2016). Intraclass Correlation Coefficient values between 0.60 and 0.74 are good and values between 0.75 and 1.0 are excellent (Cicchetti, 1994). The ICC for the STS-SRT raters was 0.900 (95% confidence

Table 3. Participants' gender, parent hearing status, Age, Age of Acquisition, and STS-SRT score

ID	Gender	Family	Age	AoA	SRT score
CHILDREN					
1611	Female	Deaf	10	0	6
1613	Female	Hearing	11	2	2
1617	Male	Hearing	11	2	1
1619	Female	Deaf	11	0	9
1620	Male	Deaf	10	0	11
1701	Female	Deaf	10	0	10
1702	Female	Deaf	11	0	14
1628	Male	Hearing	15	2	2
1629	Female	Hearing	15	1	2
1631	Female	Deaf	15	0	11
1635	Female	Hearing	15	1	5
1638	Male	Hearing	15	1	5
1647	Male	Hearing	14	7	1
1650	Male	Hearing	16	9	3
ADULTS					
s001	Female	Deaf	49	0	26
s002	Female	Hearing	26	1	24
s004	Male	Hearing	41	2	17
s008	Female	Deaf	20	0	18
s009	Male	Hearing	39	7	12
s010	Male	Hearing	38	6	8
s011	Female	Deaf	34	0	24
s013	Female	Hearing	36	2	14
s014	Male	Deaf	39	0	18
s015	Male	Deaf	23	0	21
s016	Female	Hearing	44	3	18
s017	Male	Hearing	20	1	22
s018	Female	Deaf	34	0	21
s019	Female	Hearing	39	4	6
s020	Male	Hearing	30	1	23
s021	Female	Hearing	25	2	14
s022	Female	Hearing	24	1	21

(*Continued*)

Table 3. (*Continued*)

ID	Gender	Family	Age	AoA	SRT score
s023	Female	Deaf	28	0	19
s024	Female	Deaf	43	0	20
s025	Male	Hearing	31	1	14
s026	Female	Hearing	23	2	16
s027	Female	Hearing	23	0	23
s028	Female	Hearing	22	2	21
s029	Male	Hearing	31	1	11
s030	Female	Deaf	38	0	17
s032	Female	Deaf	33	0	23
s034	Male	Deaf	33	0	17
s035	Male	Hearing	43	2	16
s036	Male	Hearing	31	2	12
s039	Male	Hearing	27	1	14

Note: AoA = Age of Acquisition.

interval $[CI] = .787, .953, p < .001$), which is in the excellent range of inter-rater reliability. Cronbach's alpha was used to determine the STS-SRT's internal reliability. Analyses revealed that the test has an alpha of 0.915 (95% $CI = .875, .948, p < .001$), suggesting that there is excellent internal consistency between the items. As an additional analysis and comparison, internal consistency was measured for the three sentence sources, Translated ($n = 11$), Corpus ($n = 8$), and Novel ($n = 12$), yielding Cronbach's alpha of 0.774, 0.714, and 0.828, and 95% $CI = [.661, .862], p < .001, [.565, .826], p < .001$, and $[.742, .894], p < .001$, respectively.

Validity evidence

Validity support for the claim that the STS-SRT results can be interpreted as a deaf test taker's language competency in STS can be provided by demonstrating that those with greater language mastery (i.e., adults or those with exposure to the language from birth) perform better on the test than those who are expected to have lesser language mastery (i.e., children or those who have experienced language deprivation). There were some factors that were introduced to this test such as the sentences originating from three different sources (ASL, STS, Corpus) and the sentences varied in the Number of Signs (2–4, 5–6, and 7+). A linear mixed model, with Modified Forward Selection, was used to capture the fixed main effects and interaction between Source and Number of Signs, with response variable the percent correct within each Source/Number of Signs combination. The independent variable contains a mix of random effects and fixed effects. Participant ID was used to account for the multiple measurements on each participant, and was a factor with random effects. Source, Number of Signs, Child_Adult (Child, Adult) Family (Deaf

Table 4. Participants' STS-SRT raw scores by children or adult and parental hearing status

Group	n	Mean	SD
Adult	30	17.67	4.93
Children	14	5.86	4.37
Deaf parents	17	16.76	5.74
Hearing parents	27	12.11	7.67

Table 5. Participants' STS-SRT raw scores by Age of Acquisition Group

AoA Group	n	Mean	SD
Birth	18	17.11	5.56
1 year old	10	14.10	8.23
2 years old	10	11.50	7.18
3+ years old	6	8.00	6.23

Note: AoA = Age of Acquisition.

family, Hearing family), and AoA Group (0, 1, 2, or 3+ years) were the factors with fixed effects, and Age, AoA, and Years of Signing were the covariates. See Tables 4 and 5 for raw scores and Table 6 for mean percent correct repetitions.

The Mixed-Effects ANOVA Analysis in Minitab was able to run the full model by using a restricted maximum likelihood method of variance estimation. However, this model suffered from multicollinearity. Therefore, the Akaike Information Criterion (AIC, corrected for small sample size) was used to build the appropriate model in order to evaluate the effects of a base set of variables related to STS-SRT performance, while accounting for subject variance and avoiding multicollinearity. Modified Forward Selection is a model-building technique in which a base set of variables (Source, Number of Signs, Source × Number of Signs) appear in each model, and additional variables are considered one a time, starting with Child_Adult, Family, Age, AoA, AoA_Group, and Years_Signing in Model 1. Child_Adult had the lowest AICc (92.24) and, hence, was added to the base set of variables for Round 2. In Round 2, among all potential models at most one additional variable over the base set of variables, the model with AoA_Group had the lowest IACc (80.64) and was added to the base set of variables for Round 3. In Round 3, the model with no additional variables had the lowest AICs (80.64), so Model 3 was used in the following analyses to determine the main effects and coefficients of the variables.

With a −2 log likelihood of 76.612, it was found that ID accounted for 9.93% of the estimated subjects' variance ($z = 2.169$, $p = .015$), leaving a substantial amount of between-subject variability, 90.07% ($z = 13.115$, $p < .001$). Main effects were found for all fixed effects: Source $F(2, 344) = 3.07$, $p = .048$, $\eta_p^2 = 0.018$; Number of Signs $F(2, 344) = 142.80$, $p < .001$, $\eta_p^2 = 0.454$; Source × Number of Signs $F(4, 344) = 4.61$, $p < .001$, $\eta_p^2 = 0.051$; Child_Adult $F(1, 39) = 126.33$, $p < .001$,

Table 6. Percent correct repetitions based on sentence length by family and age group

Sentence length	Family	Age group	Mean % Correct	SD	n
3–4 signs	Deaf parents	Adult	89	12	11
		Child	54	13	6
		Total	77	21	17
	Hearing parents	Adult	81	16	19
		Child	23	12	8
		Total	64	30	27
	Total	Adult	84	15	30
		Child	37	20	14
		Total	69	28	44
5–6 signs	Deaf parents	Adult	78	12	11
		Child	40	11	6
		Total	64	22	17
	Hearing parents	Adult	60	24	19
		Child	58	7	8
		Total	44	32	27
	Total	Adult	67	22	30
		Child	20	19	14
		Total	52	30	44
7+ signs	Deaf parents	Adult	36	18	11
		Child	8	15	6
		Total	26	22	17
	Hearing parents	Adult	23	16	19
		Child	0	0	8
		Total	16	17	27
	Total	Adult	28	18	30
		Child	3	10	14
		Total	20	19	44

$\eta_p^2 = 0.764$; and, AoA_Group $F(3, 39) = 12.11$, $p < .001$, $\eta_p^2 = 0.482$. Model 3 accounted for 65.47% of the residual variance (S = 0.236, AIC = 80.64, BIC = 88.51). See Table 7 for a summary of the results. Partial eta-squared is the variance explained by a given variable of the variance remaining after excluding variance explained by other predictors.

Adults scored 20.2% higher than children, and earlier AoA was associated with higher scores. With the base level for AoA Group being 3+ years, participants who

Table 7. Coefficients for percent correct STS-SRT repetitions

Variable	Estimate	SE	95% onfidence Interval		t-test (df)	p	Effect size (d)	Reference levels
			Lower	Upper				
Source								
ASL	0.030	0.017	−0.003	0.063	1.810 (344)	0.071	0.222	vs Novel
Corpus	−0.040	0.017	−0.073	−0.007	−2.371 (344)	0.018	0.291	vs Novel
Number of Signs								
3–4	0.220	0.017	0.187	0.253	13.119 (344)	0.000	1.614	vs 7andUp
5–6	0.045	0.017	0.018	0.078	2.667 (344)	0.008	0.328	vs 7andUp
Source × Number of Signs								
ASL 3–4	0.013	0.024	−0.034	0.059	0.528 (344)	0.598	0.112	vs STS 7andUp
ASL 5–6	0.052	0.024	0.005	0.098	2.176 (344)	0.030	0.464	vs STS 7andUp
Corpus 3–4	0.052	0.024	0.006	0.099	2.208 (344)	0.028	0.470	vs STS 7andUp
Corpus 5–6	−0.027	0.024	−0.076	0.017	−1.248 (344)	0.213	0.266	vs STS 7andUp
Child or Adult	0.202	0.018	0.166	0.239	11.240 (39)	0.000	1.212	Adult versus Child
AoA Group								
Birth	0.144	0.026	0.091	0.196	5.556 (39)	0.000	0.873	vs 3+ years
1 year old	0.034	0.031	−0.029	0.096	1.095 (39)	0.280	0.188	vs 3+ years
2 years old	−0.037	0.031	−0.099	0.025	−1.195 (39)	0.239	0.205	vs 3+ years

were exposed to STS from birth performed 14.4% better than those who acquired STS when they were 3 years old or older. Those who acquired STS around the age of 1 or 2 years old did not have significantly more correct repetitions than those who acquired the language later. With the base level for Source being STS, participants performed 3.0% better on the sentences that were translated from the ASL-SRT than on the STS sentences that the team had developed. However, correct reproduction was 4.0% less on the sentences developed from the STS corpus than on items developed by the team. The results also indicated that participants performed 22.0% better on items of 3–4 signs than on items of 7+ signs, but only 4.5% better on items of 5–6 signs.

Discussion

Similar to the ASL-SRT (Hauser et al., 2008), the STS-SRT finds that adults produce more correct repetitions than children, and those with early AoA produce more correct repetitions than those with delayed STS acquisition. The authors utilize the *argument-based validation framework* (Chapelle, 2020; Knoch & Chapelle, 2018) to describe the validity of STS-SRT interpretation. This framework requires test developers to be explicit about the claims and inferences they make about the test, and about its interpretation and use. Here, we make *evaluation inferences* (Kane, 2006) that the STS-SRT is a good measure of STS proficiency both because of how it was developed and because of its results. The sentences were carefully crafted by qualified individuals with extensive STS linguistic knowledge, and demonstrated excellent internal consistency. The scoring protocol supported this with excellent inter-rater reliability. The inference that the test is good was further supported by results demonstrating that the translated, corpus, and novel sentences did not impact sentence repetition differently. This is an *explanation inference,* about why the test scores represent an individual's language proficiency. It supports past research suggesting that the SRT measures language, not memory *per se*. This inference was supported by results demonstrating that native signers perform better than non-native signers and that adults perform better than children. The sign span results also provide evidence to support explanation inferences about the results because this study demonstrated that longer sentences cannot be reproduced with rote memory alone, providing support that implicit linguistic knowledge is needed to correctly repeat sentences.

Age of Acquisition accounted for 14.4% of the variance of the STS-SRT participants' correct repetitions. It is unclear what explains the remaining variance, although in the USA, Stone et al. (2015) found that AoA explained only 15.2% of the variance in ASL-SRT correct repetitions. Our hypothesis is that there are fewer individuals in Sweden who experience language acquisition delays compared to in other countries because Sweden has a relatively long history with an infrastructure of sign language intervention for hearing parents. This intervention includes teaching STS to hearing parents of deaf children during the children's first years. Many deaf children of hearing parents have therefore been exposed to STS from early on and have access to STS at home. As can be seen in the AoA distribution of the participants of this study, there are relatively few participants who have had first exposure to STS after the age of 3. In fact, most deaf children will have been exposed to STS through their entire childhood up to adult years. Nevertheless, it is hard to track the degree of exposure, and one would expect more variability in STS proficiency within that group compared to among deaf children of deaf parents, who are expected to have been exposed to STS from birth.

One limitation of the study is the small sample size. There is no official record of the size of the deaf (since childhood) population in Sweden, but one recurring estimate is 8,000–10,000 people in total (Parkvall, 2015). The deaf population is thus much smaller than in some other countries such as the USA. This makes it difficult to have a large sample size of subgroups within the deaf community such as native signers, who typically represent only 5–10% of the deaf community (Mitchell & Karchmer, 2004). Regardless, no test should be considered valid based on a single

psychometric property or publication. Under the argument-based validity framework, we can justify the claim that STS-SRT scores can reveal how well a person is able to comprehend and reproduce STS sentences compared to others. However, more studies are needed to add to the overall construct validity of the STS-SRT as a test of STS proficiency. Future studies of the STS-SRT are needed to determine, for example, test–retest reliability and how claims and inferences about the test's interpretation and use are validated.

Another possible limitation pertains to the test scoring system itself. There is rigorous research on the importance of nonmanual components for the structure of sign languages, as well as for the emergence of the same in stages of sign language development. However, ASL-SRT inter-rater reliability was poor when the test included the scoring of nonmanual components. The gradual variation of nonmanual components among the signers was highly variable, which made it hard for the raters to detect and rate them. STS-SRT followed the same scoring protocol as ASL-SRT, but added a limited set of nonmanual components in the scoring protocol, namely mouthing and headshaking for grammatical functions, in order to increase the efficiency of rating. It is advisable, however, that future tests include the scoring of the nonmanual components in order to capture sign language proficiency more holistically.

Most importantly, the results of this study support the Critical Period Hypothesis for language acquisition (see Newport et al., 2001 for discussion). The compromised performance by deaf individuals with delayed language exposure, and/or those raised in inaccessible, language-impoverished environments, is a concern because language deprivation is prevalent in the deaf community (Hall, 2017; Mayberry, 2010; Mayberry, Lock & Kazmi, 2002). Research has demonstrated that the early acquisition of sign language is important for cognitive (e.g., Courtin, 2000; Hall et al., 2018; Hauser, Lukomski & Hillman, 2008) and literacy development (e.g., Clark et al., 2016; Stone et al., 2015). The results of this study provide evidence not only of the validity of the STS-SRT, but also that delayed sign language acquisition is happening in Sweden as it is elsewhere.

Conclusion

This study on the one hand provides straightforward implications for sign language testing and on the other hand, widens our understanding of language testing more generally. This study proposes an STS test that has excellent reliability and validity, and which can be used by researchers and practitioners to document child and adult STS proficiency. The STS-SRT can be used as a continuous variable in theoretical, clinical, and applied psycholinguistics and other language and behavioral sciences, as well as by educators and other specialists tracking language development.

earlier drafts of this paper. Finally, we would like to thank our incredible test takers for participating in the study.

References

Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading*, 33, 206–219. https://doi.org/10.1111/j.1467-9817.2009.01412.x

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, 52, 899–914. https://doi.org/10.1044/1092-4388(2009/08-0099)

Bogliotti, C., Aksen, H., & Isel, F. (2020). Language experience in LSF development: Behavioral evidence form a sentence repetition task. *PLoS ONE*, 15, e0236729. https://doi.org/10.1371/journal.pone.0236729

Brownell, C. A. (1988). Combinatorial skills: Converging developments over the second year. *Child Development*, 59, 675–685. https://doi.org/10.2307/1130567

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: Sage Publications.

Cicchetti, C. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clark, D. M., Hauser, P. C., Miller, P., Kargin, T., Rathmann, C., Guldenoglu, B., Kubus, O., & Israel, E. (2016). The importance of early sign language acquisition for Deaf readers. *Reading and Writing Quarterly*, 32, 127–151. https://doi.org/10.1080/10573569.2013.878123

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, 42(6), 741–748. https://doi.org/10.1111/1469-7610.00770

Cormier, K., Adam, R., Rowley, K., Woll, B., & Atkinson, J. (2012, March). The British Sign Language Sentence Reproduction Test: Exploring age-of-acquisition effects in British deaf adults. Paper presented at "Experimental Studies in Sign Language Research: Sign Language Workshop" at the *Annual Meeting of the German Linguistics Society* (DGfS), Frankfurt am Main, Germany.

Coughlin, C. E., & Tremblay, A. (2013). Proficiency and working memory based explanations for nonnative speakers sensitivity to agreement in sentence processing. *Applied Psycholinguistics*, 34, 615–646. https://doi.org/10.1017/S0142716411000890

Courtin, C. (2000). The impact of sign language on the cognitive development of deaf children: The case of theories of mind. *Journal of Deaf Studies and Deaf Education*, 5(3), 266–276. https://doi.org/10.1093/deafed/5.3.266

Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language and Communication Disorders*, 42, 187–208. https://doi.org/10.1080/13682820601030686

Enns, C., Haug, T., Herman, R., Hoffmeister, R., Mann, W., & McQuarrie, L. (2016). Exploring signed language assessment tools in Europe and North America. In M. Marschark, V. Lampropoulou, & E. Skordilis (Eds.), *Diversity in Deaf Education* (pp. 171–218). New York: Oxford University Press.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27, 464–491. https://doi.org/10.1093/applin/aml001

Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R. (2016). How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition*, 25, 85–101. https://doi.org/10.1080/10489223.2016.1192635

Gaillard, S., & Tremblay, A. (2016). Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. *Language Learning*, 1–29. https://doi.org/10.1111/lang.12157

Hall, M. L., Eigsti, I-M., Bortfeld, H., & Lillo-Martin, D. (2018). Executive function in deaf children: Auditory access and language access. *Journal of Speech, Language, and Hearing Research*, **61**, 1970–1988. https://doi.org/10.1044/2018_JSLHR-L-17-0281

Hall, W. C. (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

Hammill, D., Brown, V., Larsen, S., & Wiederholt, J.L. (1994). *Test of Adolescent and Adult Language* (3rd Ed.). Austin, Texas: PRO-ED, Inc.

Haug, T., Batty, A.O., Venetz, M., Notter, C., Girard-Groeber, S., Knoch, U., & Audeoud, M. (2020) Validity evidence for a sentence repetition test of Swiss German Sign Language. *Language Testing*, **37**, 412–434. https://doi.org/10.1177/0265532219898382

Hauser, P. C., Lukomski, J., & Hillman, T. (2008). Development of deaf and hard of hearing students' executive function. In M. Marschark & P. C. Hauser (Eds.), *Deaf cognition: Foundations and outcomes* (pp. 286–308). New York: Oxford University Press.

Hauser, P. C., Paludnevičiene, R., Supalla, T., & Bavelier, D. (2008). American Sign Language-Sentence Reproduction Test: Development and implications. In R. M. de Quadros (ed.), *Sign Language: Spinning and unraveling the past, present and future* (pp. 160–172). Petropolis, Brazil: Editora Arara Azul. https://scholarworks.rit.edu/other/596

Henner, J., Novogrodsky, R., Reis, J., & Hoffmeister, R. (2018). Recent issues in the use of sign language assessments for diagnosis of language disorders in signing deaf and hard of hearing children. *Journal of Deaf Studies and Deaf Education*, **23**, 307–316. https://doi.org/10.1093/deafed/eny014

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D. J., Padden, C., & Rathmann, C. (2014). Ensuring language acquisition for deaf children: What linguists can do. *Language*, **90**(2), e31–e52. https://doi.org/10.1353/lan.2014.0036

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, **64**(1), 215-238. https://doi.org/10.3138/cmlr.64.1.215

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Westport, CT: American Council on Education.

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes*, **22**, 860–897. https://doi.org/10.1080/01690960601155284

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, **18**(1), 146–154. https://doi.org/10.1111/desc.12202

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, **35**(4), 477–499. https://doi.org/10.1177/0265532217710049

Koo, T. K., & Li, M. Y. (2016). A guideline for selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, **15**, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kubus, O., Villwock, A., Morford, J. P., & Rathmann, C. (2015). Word recognition in deaf readers: Cross-language activation of German Sign Language and German. *Applied Psycholinguistics*, **36**(4), 831–854. https://doi.org/10.1017/S0142716413000520

Landa, R., & Clark, M. (2019) L2/Ln sign language tests and assessment procedures and evaluation. *Psychology*, **10**(2), 181–198. https://doi.org/10.4236/psych.2019.102015

Leclercq, A-L., Quémart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A powerful diagnostic tool for French children with specific language impairment. *Research in Developmental Disabilities*, **35**, 3423–3430. https://doi.org/10.1016/j.ridd.2014.08.026

Mahshie, S.N. (1995). *Educating deaf children bilingually. With insights and applications from Sweden and Denmark*. Washington, DC: Pre-College Programs, Gallaudet University.

Mayberry, R. I. (2010). Early language acquisition and adult language ability: What sign language reveals about the critical period for language. In M. Marschark & P. E. Spencer (Eds.), *The Oxford handbook of deaf studies, language, and education* (Vol. 2, pp. 281–291). New York, NY: Oxford University Press.

Mayberry, R. I., Lock, E., & Kazmi, H. (2002). Linguistic ability and early language exposure. *Nature*, **417**(6884), 38–38. https://doi.org/10.1038/417038a

Mesch, J., & Wallin, L. (2012, May). From meaning to signs and back: Lexicography and the Swedish Sign Language Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 123–126). https://www.sign-lang.uni-hamburg.de/lrec2012/programme.html

Mitchell, R. E., & Karchmer, M. A. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies*, **4**(2), 138–163. https://doi.org/10.1093/deafed/enh017

Newport, E. L., Bavelier, D., & Neville, H. J. (2001). Critical thinking about critical periods: Perspectives on a critical period for language acquisition. In E. Doupoux (Ed.), *Language, brain and cognitive development: Essays in honor of Jacques Mehler* (pp. 481–502). Cambridge, MA: MIT Press.

Paludnevičience, R., Hauser, P. C., Daggett, D. J., & Kurz, K. B. (2012). Issues and trends in sign language assessment. In D. Morere & T. Allen (Eds.), *Measuring literacy and its neurocognitive predictors among deaf individuals: An assessment toolkit* (pp. 191–207). New York: Springer.

Parkvall, M. (2015). *Sveriges språk i siffror: Vilka språk talas och av hur många?* [Languages of Sweden in numbers. What languages are spoken and by how many?] Institutet för språk och folkminnen, Språkrådet/Morfem.

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure? *International Journal of Language and Communication Disorders*, **50,** 106–118. https://doi.org/10.1111/1460-6984.12126

Proposition 1980/81:100, Bilaga 12. Stockholm: Utbildningsdepartementet.

Quinto-Pozos, D., Singleton, J., & Hauser, P. C. (2017). A case of specific language impairment in a Deaf signer of American Sign Language. *Journal of Deaf Studies and Deaf Education*, **22,** 204–218. https://doi.org/10.1093/deafed/enw074

Rinaldi, P., Caselli, M. C., Lucioli, T., Lamano, L., & Volterra, V. (2018). Sign Language Skills Assessed Through a Sentence Reproduction Task. *The Journal of Deaf Studies and Deaf Education*, **23**(4), 408–421. https://doi.org/10.1093/deafed/eny021

Schönström, K., Hauser, P. C., & Rathmann, C. (in press). Validation of signed language tests for adult L2 learners. In T. Haug, W. Mann, & U. Knoch (Eds.), *The Handbook of Language Assessment Across Modalities*. London: Oxford University Press.

SFS 2009:600. *Språklag* [Language Act]. Stockholm: Kulturdepartementet.

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, **65**, 723–751. https://doi.org/10.1111/lang.12129

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research*, **49**, 219–236. https://doi.org/10.1044/1092-4388(2006/019)

Stone, A., Kartheiser, G., Hauser, P. C., Petitto, L-A., & Allen, T. E. (2015). Fingerspelling as a noval gateway into reading fluency in deaf bilinguals. *PLoS One*, **10**, e0139610. https://doi.org/10.1371/journal.pone.0139610

Supalla, T., Hauser, P. C., & Bavelier, D. (2014). Reproducing American Sign Language sentences: Cognitive scaffolding in working memory. *Frontiers in Psychology*, **5**, 859. https://doi.org/10.3389/fpsyg.2014.00859

Svartholm, K. (2010). Bilingual education for deaf children in Sweden. *International Journal of Bilingual Education and Bilingualism*, **13**(2), 159–174. https://doi.org/10.1080/13670050903474077

Tomblin, J.B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language and Hearing Research*, **49**(6), 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)

van den Noort, M. W. M. L., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist*, **11**, 289–296. https://doi.org/10.1027/1016-9040.11.4.289