Validity evidence for a sentence repetition test of Swiss German Sign Language
[ACCEPTED VERSION]

Tobias Haug[1], Aaron Olaf Batty[2], Martin Venetz[1], Christa Notter[3], Simone Girard-Groeber[4], Ute Knoch[5], and Mireille Audeoud[1]

[1] University of Applied Sciences of Special Needs Education, Switzerland

[2] Keio University, Japan

[3] Dima Language School, Switzerland

[4] University of Applied Sciences and Art of Northwestern Switzerland

[5] University of Melbourne, Australia

Abstract

This article reports on a study seeking evidence of validity according to the socio-cognitive framework (Weir, 2005) for a new sentence repetition test (SRT) for young Deaf L1 Swiss German Sign Language (DSGS) users. SRTs have been developed for various purposes for both spoken and sign languages to assess language development in children. In order to address the need for tests to assess the grammatical development of Deaf L1 DSGS users in a school context, an SRT was developed. The test targets young learners aged 6–17 years old, and was administered to 46 Deaf students aged 6.92–17.33 (M = 11.17) years. In addition to the young learner data, data were collected from Deaf adults (N = 14) and from a sub-sample of the children (n = 19), who also took a test of DSGS narrative comprehension, serving as a criterion measure. Data were analyzed with many-facet Rasch modeling, regression analysis, and analysis of covariance. The results show evidence of scoring, criterion, and context validity, suggesting the suitability of the SRT for the intended purpose, and will inform the revision of the test for future use as an instrument to assess the sign language development of Deaf children.

Introduction

Tests of sign language proficiency are a growing sub-field of language testing, one that is facing the same issues and challenges as in the assessment of spoken languages. Although the number of tests available for sign languages is increasing, tests to assess the grammatical development of Deaf[1] children who use Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) as their primary language have only recently begun to be developed. One method of assessment of child grammatical development that has been widely used in spoken language contexts is the sentence repetition tests (SRT). SRTs exist for other sign languages; for example, American Sign Language (ASL; Hauser, Paludnevičienė, Supalla, & Bavelier, 2008), but none for DSGS. The development of an SRT for DSGS, as for any other sign or spoken language, requires the establishment of validity evidence, which is the focus of this study.

The overall goal of the present research was to develop and evaluate a sentence repetition test (SRT) for DSGS that targets Deaf children and adolescents aged 6–17 years old. Existing SRTs for American Sign Language (Hauser et al., 2008), British Sign Language (BSL; Marshall et al., 2015), German Sign Language (*Deutsche Gebärdensprache*, DGS; Kubus & Rathmann, 2012), Italian Sign Language (*Lingua dei Segni Italiana*, LIS; Rinaldi, Caselli, Lucioli, Lamana, & Volterra, 2018), and Swedish Sign Language (*Svenskt Teckenspråk*, STS; Schönström & Holmström, 2017) were used as a framework to inform the development of the DSGS sentences, as well as studies on SRTs for spoken languages.

Literature review

Sentence repetition tests (SRTs)

Sentence repetition tests have been developed for various purposes; for example, to assess language acquisition in typically developing children (e.g., Devescovi & Caselli, 2007; Klem et al., 2015), as a tool to investigate language proficiency in adult learners of a second language (e.g., Gaillard & Tremblay, 2016; Spada, Shiu, & Tomita, 2015), or as a clinical marker to identify a specific language impairment (e.g., Conti-Ramsden, Botting, & Faragher, 2001; Meir, Walters, & Armon-Lotem, 2016; Poll, Betz, & Miller, 2010).

An underlying assumption of an SRT is "if the participant has acquired the grammatical feature associated with or displayed in the stimuli, it should be easy to repeat the stimuli" (Yan, Maeda, Lv, & Ginther, 2016, p. 498). SRTs involve (1) the processing of a stimulus sentence, (2) reconstructing it with the test-takers' own grammar, and (3) reproducing it (Jessop, Suzuki, & Tomita, 2007).

There does not seem to be a consensus on what exact construct an SRT taps into, although several researchers have attempted to address this. Yan et al. (2016) reviewed a number of studies using SRTs, concluding that the measured construct can be summarized as (1) a global proficiency or (2) more specific linguistic features, for example, phonology, morphosyntax, and syntax (p. 504). Okura and Lonsdale (2012) raise the question of whether the construct addressed by SRTs is one of language proficiency or, rather, rote repetition, whereas Spada and colleagues (2015), in a study of implicit linguistic knowledge in L2 adult learners, argue that the construct addressed is grammatical processing.

[1]It is a widely-recognized convention to use upper case *Deaf* for describing members of the linguistic community of sign language users and, in contrast, the lower case *deaf* for describing individuals with an audiological state of a hearing impairment, not all of whom might be sign language users (Morgan & Woll, 2002).

In an attempt to settle this controversy, some researchers have added ungrammatical sentences to their stimuli with the expectation that an ungrammatical sentence processed by a test-taker will result in a corrected sentence (Erlam, 2006; Sarandi, 2015; Yan et al., 2016). For example, in the study by Erlam (2006), native speakers of English ($N = 20$) corrected 91% of ungrammatical sentences in an SRT (and repeated 97% of grammatical sentences correctly) which the author interprets as "evidence of the validity of the test as a measure of implicit [linguistic] knowledge" (p. 485). Additional evidence that the measured construct is linguistic knowledge is a study by Klem et al. (2015). Klem and colleagues (2015) investigated an SRT as a measure of language ability in school-aged children ($N = 216$) in the Norwegian context, concluding, "sentence repetition is best conceptualized as a measure of language ability" (p. 152). The authors further argue that "sentence repetition is best seen as a complex linguistic task that reflects the integrity of language processing systems at many different levels (speech perception, lexical (vocabulary) knowledge, grammatical skills, and speech production […])" (p. 152).

Support for the notion that the SRT format measures linguistic knowledge has been further provided by various studies (e.g., Devescovi & Caselli, 2007; Graham et al., 2010; Jones, 1994). For example, Devescovi and Caselli (2007) used an SRT for spoken Italian with pre-schoolers aged 2–4 years old ($N = 25$) and compared the results with the children's spontaneous language data. There were significant positive correlations between the mean length of utterance, omission of articles, and number of verbs produced in both measures. The authors conclude that an SRT can be used (along with other measures) "to evaluate language abilities in typical developing children between 2 and 4 years of age." (Devescovi & Caselli, 2007, p. 201).

SRTs for sign languages

Only a few studies published in the literature explicitly address the development of SRTs for sign languages. For example, Hauser et al. (2008) discuss the development of an SRT for American Sign Language (ASL) as a global measure of proficiency to test Deaf and hearing signers at different levels. The ASL SRT was used both with adults and children (age range children: 12.5 to 14.1 years old; $M_{age} = 12.9$). The ASL SRT is based on the *Speaking Grammar Subtest* of the *Test of Adolescent and Adult Language – Third Edition* (TOALT3; Hammill, Brown, Larsen, & Wiederholt, 1994). In total, 40 sentences in increasing length and of different syntactic, thematic, and morphological complexity were developed (Hauser et al., 2008). Difficulty was increased, for example, by using more complex morphological signs. It was found that increasing sentence length did not automatically increase the complexity of a sentence in sign languages (Hauser et al., 2008). The ASL SRT has also been adapted to German Sign Language (*Deutsche Gebärdensprache*; DGS) (Kubus & Rathmann, 2012), British Sign Language (BSL; Cormier, Adam, Rowley, Woll, & Atkinson, 2012), and Swedish Sign Language (*Svenskt Teckenspråk*, STS) (Schönström & Holmström, 2017).

Socio-cognitive approach to test validation

The socio-cognitive approach to test validation (O'Sullivan & Weir, 2011; Weir, 2005) includes the cognitive, social, and evaluative dimension of "language use in test development and validation" (O'Sullivan & Weir, 2011, p. 20). This approach includes various validity arguments for which evidence is gathered at different stages of test development and use, collectively contributing to an argument for the overall validity of the test. The kinds of validity evidence are: (1) test-takers' characteristics, (2) context validity (i.e., characteristics of test tasks and their administration), (3) cognitive validity (i.e., appropriateness of cognitive processes required to complete the tasks), (4) scoring validity (i.e., meaning of the score), (5) consequential validity (i.e., effect of test in stakeholders), and (6) criterion-related validity (i.e., other/external evidence than the test scores showing that test is doing a good job) (O'Sullivan

&Weir, 2011; Weir, 2005). This framework will serve as the basis for the present validation of the SRT for DSGS, with particular attention paid to test-taker characteristics, scoring and criterion-related validity. Additionally, context validity was partly addressed in item and rating scale development.

Test-takers characteristics

Test-taker characteristics such as chronological age or parental hearing status have often been used as a means to differentiate between early or later access to a sign language (L1 vs. L2) in studies evaluating sign language tests (e.g., Herman, 2002; Mann, 2006). The variable of chronological age is used to investigate whether a test instrument represents developmental progression in signing Deaf children (Herman, 2002). The variable of parental hearing status is often used to account for the heterogenous linguistic experiences of Deaf children (Mayberry, Lock, & Kazmi, 2002). Only about 5% of Deaf children are born into Deaf families and therefore may have access to a sign language from birth as a first language. The remaining 95% are born into non-signing hearing families (Mitchell & Karchmer, 2004) and might have first access to sign language after the critical period of language acquisition (Mayberry et al., 2002).

A group of native signing Deaf children is therefore often used as a model or a reference against which the performances of children with other linguistic experiences (Deaf children of hearing parents) can be measured. It is important to point out, however, that the use of parents' hearing status as a variable is not entirely undisputed, as Deaf parents may not be native signers *per se*, as they may have grown up in a hearing family and learned sign language later (e.g., Singleton & Newport, 2004). For the purpose of determining whether the DSGS SRT scores align with developmental expectations, both variables will be included in the model of evaluating the SRT for DSGS.

Deaf adults as a reference for test development

Many sign languages are not as well researched as spoken languages. This at least partly accounts for the incomplete description of DSGS grammar as well as the lack of L1 DSGS child acquisition studies to use as external "benchmarks" to inform the development of items for the measurement of developmental progression. As a result, Deaf adults' performances are the only practical point of reference for mastered/acquired structures of DSGS. Since the present test is the first for DSGS targeting Deaf children, the performances of Deaf adults were compared with those of the target population of children and adolescents (e.g., Rinaldi et al., 2018), on the assumption that adult users of DSGS should outperform the children since the adults had already acquired DSGS fully.

Linguistic structure of DSGS

Linguistic structures of DSGS that are part of the construct of the SRT will be briefly described in this section. An important feature in any sign language is the distinction between *manual* and *non-manual* components. Manual components are produced with the hands; non-manual components are features that are produced with the mouth, the face (e.g., with cheeks, eyes, eyebrows, etc.), with the head, and the upper torso (Boyes Braem, 1995; Sutton-Spence & Woll, 1999). For example, eye gaze can be used to re-establish reference in signing space or raised eyebrows to differentiate between a declarative and an interrogative sentence (Pfau & Quer, 2010).

Another important feature of sign languages is the use of *signing space*, i.e., the physical space in front of the signer's body, which serves various purposes (Johnston & Schembri, 2007). The signing space is important for introducing and maintaining reference. For example, with the first mention of an object or person an index is used to locate it (e.g., a house) at a specific point in space. With gaze or index finger at this same locus the signer can then establish

pronominal reference (Boyes Braem 1995). The signing space is also important in representing how an object (e.g., a car) moves from A to B.

Sign language phonology: The smallest building blocks of sign languages are the sub-lexical units of signs. These sub-lexical units are the handshape, location, movement, and hand orientation (Boyes Braem, 1995).

Sign language morphology: Another important aspect of sign language morphology is verb classes, which are, depending on the underlying model, grouped as *plain*, *agreement*, and *spatial verbs* (e.g., Padden, 1990). Another area is *negation* which is expressed in DSGS manually, non-manually, or by a combination of both.

Sign language syntax: Sign languages are described as having more flexible word/sign order than spoken languages (Erlenkamp, 2012). The difference between a question and a statement is expressed non-manually. For example, the sign GEHÖRLOS [2] (deaf) in a declarative sentence "I am deaf" shows a neutral facial expression and head position. This is different if the sign is part of a question such as "Are you deaf?". Here, the sign is realized with a slight head movement forward and raised eyebrows (Boyes Braem, 1995).

Discourse strategies: A frequently used discourse strategy in sign languages is *constructed action*, (e.g., for BSL: Cormier, Smith, & Zwets, 2013). Constructed action refers to a situation in which the signer "takes the role" of a referent to express his or her feelings, ideas, actions, etc. The signer uses manual and non-manual techniques to express specific feelings or actions of a referent.

Research questions

The present study seeks validity evidence for a new sentence repetition test (SRT) for school-aged Deaf users of Swiss German Sign Language (DSGS) through the following research questions (RQs):

RQ1: To what extent does the DSGS SRT demonstrate evidence of scoring validity?

RQ2: To what extent does the DSGS SRT demonstrate evidence of criterion-related validity?

RQ3: To what extent do individual test-taker characteristics (age and hearing status of parents) impact performance on the DSGS SRT?

RQ4: To what extent does the DSGS SRT demonstrate evidence of context validity?

Method

Instruments

DSGS Sentence Repetition Test (SRT)

Existing SRTs for sign languages described above were used as a framework for developing the SRT for DSGS, referring both to the sentences and to the scoring criteria, to be detailed below.

SRT item development

The content of the SRT for the current study was developed through a process of expert moderation. In the first step, an item candidate pool of 75 sentences was developed: 38 were

[2] GEHÖRLOS (deaf) is an example of a sign language gloss, a label for one aspect of the meaning of a sign. Glosses are typically written in all caps (Ebling, 2016). In this paper, signs from DSGS are glossed in German and the English meaning is added in parentheses.

based on the DGS version, 17 from a BSL SRT for children, and 10 from the Italian Sign Language SRT for children. The sentence development was supplemented by five sentences from the DSGS online learning materials for families with Deaf children "E-Kids" (SGB-FSS, n.d., https://ekids.sgb-fss.ch/) and five sentences that were developed by the Deaf research collaborator of the project. Even though the majority of the sentences of the SRT for DSGS are from the DGS version (which, in turn which, was a direct translation from the ASL SRT), sentences were also adapted from other SRTs developed explicitly for younger children (under the age of 12, which is the youngest age group of the SRT for ASL). The goal at this stage of the project was to have a pool of DSGS sentences available that (1) varied in length, (2) varied in complexity, and (3) were sensitive to the life experiences of children 6–17 years old. This pool was subjected to expert moderation twice by two separate panels of Deaf sign language instructors each before it was administered to any test-takers. The first panel consisted of four Deaf sign language instructors, and the second, of five. In the first moderation, the sentences were evaluated for regional variation (Haug, 2011; Hauser et al., 2008); grammaticality; and relevance to the child sample, in terms of life experience and linguistic development. This resulted in the removal of fifteen (15) sentences.

In the second moderation, the judges individually rated the sentences' difficulty from the perspective of a Deaf child on a four-point holistic Likert scale ranging from "very easy" to "very difficult." Sentences for which the judges showed little or no agreement were later discussed as a group. These discussions resulted in the following criteria for describing/separating easier from more difficult sentences: (1) length of the sentence, (2) use of non-manual components, and (3) use of space. This process of ensuring that the test tasks matched the test-takers in terms of appropriacy of information and content, grammatical and lexical difficulty, and regional language variation suggests support of a claim to context validity (O'Sullivan & Weir, 2011; Weir, 2005).

Thirty-six (36) sentences upon which the majority of the five judges of the second moderation agreed within one point on the Likert scale were kept. The remaining 39 sentences were further discussed by the Deaf research collaborator and one other Deaf colleague with extensive experience in sign language research. On the basis of this discussion, 15 of these 39 sentences were removed, resulting in a pool of 60 sentences.

The pool of 60 sentences were then piloted with a small representative sample of three signing children and two adults, and the responses utilized for rater training (described below). Based on rater recommendations from this pilot, a further 20 sentences were removed for either being too difficult (no participants repeated correctly) or too easy (all repeated correctly), resulting in a final instrument comprised of 40 sentences which had passed all three moderation/training steps. The final 40 sentences included specific linguistic features of DSGS, for example, phonology (e.g., sub-lexical units of signs), morphology (e.g., types of verbs, different forms of negation), syntax (e.g., different types of sentences), discourse strategies (e.g., constructed action), and non-manuals (e.g., negation) (e.g., Boyes Braem, 1995; Sutton-Spence & Woll, 1999).

SRT rating scale and rater training

The rating scale for the SRT was developed based on a study of three candidate scales (Batty & Haug, forthcoming) with the aim of developing a scale that offered more detailed information about the children's performances than would result from a simple dichotomous scale (Leclercq et al., 2014). The results of this study, as well as the criteria laid out by Marshall et al. (2015), informed the development of the current rating scale. The final, five-step (0–4) scale is presented in Table 1.

Table 1

Rating Scale for the Current Study

Criterion 1: Correct repetition of sentence	Criterion 2: Correct repetition of the signs (though not right order)	Criterion 3: Acceptable sign order (whole sentence)	Criterion 4: Meaning of signs is correct (not same as target)	Score	Explanation
•	•	•	•	4	Correct repetition of the target sentence (Criterion 1) includes Criteria 2–4
○	•	•	•	3	Not exact repetition, but Criteria 2–4 are correct
○	•	•	○	2	Any combination of **two** of the Criteria 2–4
○	○	•	•		
○	•	○	•		
○	•	○	○	1	**One** of the Criteria 2–4
○	○	•	○		
○	○	○	•		
○	○	○	○	0	None of the criteria

• = criterion is met; ○ = criterion is not met

Criteria 1, 2, and 4 were judged at the single sign level; only Criterion 3 was judged at the sentence level. When a single sign was judged as being incorrect (e.g., one out of four signs in a sentence), no points were assigned to the criterion. A total score of 4 was possible for each sentence, one for each criterion. When something was incorrect, it was possible to specify the error for each individual sign, for example, wrong use of non-manual features in a negated utterance or wrong sub-lexical units (e.g., incorrect handshape). However, this information did not have an impact on the test-taker's score. Non-manual features (e.g., facial expression of negation or questioning) were not listed as a separate criterion, but included in the different criteria, for example, in Criterion 3, the use of facial expression for asking a question needs to be present for the child to receive a score.

Since the goal of this rating scale was to obtain detailed information about the Deaf children's DSGS performances, rater training was also required. Rater training was conducted by the Deaf research collaborator with the two Deaf raters. The training included familiarizing the raters with the rating scale and analyzing the data from three native signing children from the item development pilot described above, including feedback and discussion on the rating criteria moderated by the Deaf collaborator. Ensuring that marking criteria are explicit for the raters further supports a claim of context validity (O'Sullivan & Weir, 2011; Weir, 2005). Furthermore, the interaction between the rating system and the scores produced, which will be discussed in greater detail with respect to the many-facet Rasch model, will provide evidence of scoring validity.

DSGS Narrative Comprehension Test

Due to the absence of another DSGS test that covers the same construct, the results of a DSGS narrative comprehension test (Haug & Perrollaz, 2015) were correlated with the SRT results in order to investigate criterion-related validity (RQ2). The narrative comprehension test was developed within the EU project SignMET (Sign Language: Methodologies and Evaluation Tools). The test was evaluated as part of the scientific final report for the funder of the project (SignMET Consortium, 2016). A total of 34 Deaf children took this test, their age ranged from 4.0 to 14.0 years of age (M_{age} = 8.67). Of these 34 children, 26 had hearing parents; the remaining 8 had at least one Deaf parent. The maximum possible score on the test was 17, and the raw scores of the children ranged from 0 to 16 ($M_{raw\ scores}$ = 9.65, $SD_{raw\ scores}$ = 5.08).

In order to investigate the relationship between chronological age, raw scores, and hearing status of the parents, a one-way, between-groups analysis of covariance (ANCOVA) was computed with the raw score as the dependent variable, the parents' hearing status the independent variable, and chronological age, the covariate. There was a statistically significant difference in the raw scores between Deaf children with hearing and Deaf parents [F (1, 31) = 6.13, p = .019, partial η^2 = .165]. The parental hearing status explains only 16.5% of the variance in the raw scores. There was also a significant relationship between the chronological age covariate and the dependent variable while controlling for the independent variable [F (1, 31) = 37.97, p < .001, partial η^2 = .551). The chronological age explains 55.1% of the variance of the raw scores.

A sub-sample (n = 19) that took part in the SRT project was also tested with this narrative comprehension test (Table 2). Even though the narrative comprehension test does not tap into exactly the same construct of the SRT, there are some aspects in the construct of the narrative comprehension test that are similar. First of all, both the SRT and the narrative test assess (also) comprehension skills. Additionally, some grammatical features are shared by the construct of both tests, for example:

(1) Signing space (e.g., for pronominal referencing)
(2) Verb type (e.g., agreement and spatial verbs)
(3) Constructed action
(4) Non-manual features for grammatical purposes (e.g., negation, asking questions)

We therefore argue, based on the preliminary statistical results and the overlap of the construct of both tests, that this comparison can be used to investigate criterion-related validity for the SRT for DSGS, thereby addressing RQ2.

Table 2
Description of the Sub-Sample of the Narrative Comprehension Test (n = *19*)

Parents' hearing status	Male	Female	Age range	M_{age}	SD_{age}
Deaf (n = 6)	5	1	7.25–13.25	9.75	2.25
Hearing (n = 13)	9	4	6.92–14.33	9.67	2.58
Total	14	5	6.92–14.33	9.50	2.42

Participants

In total 46 children and adolescents were recruited through the five schools for the Deaf in German Switzerland. They were tested between June and November 2014. Demographic

data collected included the hearing status of the participants' parents, as this is often used in sign language research as an indication of L1/L2 status (i.e., for those with Deaf parents, sign language is their L1). See Table 3 for a breakdown of participant characteristics.

Table 3
*Description of the Sample of the Main Study (*N = *46*)

Parents' hearing status	Male	Female	Age range*	M_{age}	SD_{age}
Deaf (*n* = 11)	8	3	7.33–13.33	10.25	2.00
Hearing (*n* = 35)	20	15	6.92–17.33	11.42	3.17
Total	28	18	6.92–17.33	11.17	2.83

*One missing value for the variable Age

Procedure

The entire test was embedded in a PowerPoint presentation, which was presented to the children individually on a laptop. After the pre-recorded test instructions, the children saw six practice items to become familiar with the task, followed by the 40 sentences. During the testing session, a Deaf test administrator was present and guided the children through the test. The children were video-recorded through the built-in webcam of the laptop. The testing took between 20-30 minutes. Apart from the tests, the parents filled out a background questionnaire. Parents also received background information about the study and signed a consent form. All materials were collected through the schools and returned to the researchers.

After the data collection, the video files were imported into a bespoke application for the scoring of the SRT results and given to two Deaf raters. The Deaf collaborator produced a written and a signed version of a manual, including how to use the stand-alone application of the rating scale for the raters, and also conducted a live training with them.

Due to resource constraints, it was not possible to ensure that both raters rated all children (*N* = 46). Rater 1 scored 38 children, and Rater 2 rated 22 children, with an overlap of 13 children to investigate inter-rater reliability and estimate measures in the Rasch model. This resulted in 25 cases that were evaluated only by Rater 1 and nine that were scored only by Rater 2. It took about one hour for the raters to evaluate the 40 sentences per child.

Comparing the Deaf children's and adolescents' data with results from Deaf adult signers

Data were also collected from adult Deaf signers to compare to the children's results. For this purpose, 14 Deaf adults, both "L1" and "L2" users of DSGS as defined by their parents' hearing status (i.e., Deaf vs. hearing parents), were tested with the same set of items of the SRT for DSGS (Table 4).

Table 4
Description of the Adult Sample (N = *14*)

Parents' hearing status	Male	Female	Age range*	M_{age}	SD_{age}
Deaf (*n* = 8)	3	5	28–41	33.88	5.14
Hearing (*n* = 6)	1	5	22–37	33.00	5.51
Total	4	10	22–41	33.50	5.11

*Variable Age was reported in years only by the adults

The Deaf adults filled out a background questionnaire and signed a consent form before they took the DSGS test. Rater 1 rated eight adults, and the remaining six adults were scored by Rater 2. Rater 1 and 2 were the same persons as in the main study. Despite the lack of overlap, however, severity was estimated with the ratings of the child sample (see below). These data were used to ensure that the lexicogramatical level of the SRT was appropriate for the developmental level of the target test takers (Weir, 2005), thereby – along with the process of item and rating scale development – addressing RQ4.

Data analysis

In order to investigate the four research questions, the following statistical procedures were employed.

Many-facet Rasch measurement

Many-facet Rasch measurement (MFRM; Linacre, 1994) with the software package Facets (Linacre, 2018) was employed to address RQ1 and to detect possible threats to scoring validity. This method has frequently been employed to detect and investigate rater effects (Bachman, 2004; Myford & Wolfe, 2003), but can be used wherever two aspects (facets) of a test or testing situation are thought to interact (Batty, 2014; Brunfaut, Harding, & Batty, 2018; Engelhard, 2009). In addition, Rasch residuals-based fit statistics can be used to identify poorly-performing items or raters requiring further examination or removal. Although there are no theoretical cut-off values at which an element can be considered too "noisy" to be useful, a commonly-used guideline is that offered by Wright and Linacre (1994), which considers elements with fit statistics above 2.0 as distorting or degrading measurement.

The Facets software package also provides "fair average" scores for all elements. These are provided in the original units of measurement (a five-step scale from 0 to 4 here), and represent each examinee's score, given the severity of the rater(s) the examinee was rated by. The fair average represents the score the examinee could be expected to receive, had he/she sat the test with a theoretical average-severity rater. These fair average scores will be used to investigate the impact of individual differences on scores.

The present research employs a four-facet MFRM model to investigate instrument reliability, inter-rater reliability, and to compare performance between child and adult examinees in order to demonstrate construct validity. The facets are as follows:

1. Test-takers
2. Child/Adult (dummied)
3. Rater
4. Item

The second facet (Child/Adult) is a dummy facet, not used for estimation, but is used for investigating item difficulties for the two sub samples.

In order to address RQ4 and compare the children's performances to those of the separate sample of Deaf adult sign language users, an anchored model was used. The model was first estimated using only the children ($n = 46$) and the estimates were anchored. The adult sample ($n = 14$) was then added and the model was estimated again. This ensured that the adults' level did not contribute to the calibration of the model, and that their abilities were estimated only in terms of those of the child sample.

Two initial estimations of the model revealed six items (Items 2, 14, 30, 33, 35, and 38) with Infit mean-square (MS) values exceeding 2.0, which, according to Wright and Linacre (1994) may have degraded measurement. These items were therefore removed from the model. As such, the final count of items used in estimation was reduced from 40 to 34.

Comparative analyses

To address RQ2, evidence for criterion-related validity was sought through various comparative analyses. In order to investigate the relationship between the results of the SRT and the scores on the narrative comprehension test, a Pearson product-moment correlation was calculated between the Rasch fair average SRT scores and the raw scores of the narrative test.

Additionally, to address RQ3 and determine the degree to which the test results align with factors explaining sign language acquisition, external variables were set in relation to the test results to explain performance differences (Haug, 2011; Mann, 2006). Variables that were examined were (1) chronological age and (2) hearing status of the parents, with the assumption that both age and having at least one Deaf parent would be predictive of higher scores on the SRT. In order to investigate the variable age, a simple linear regression analysis was applied. In order to investigate if the parental hearing status contributed to SRT performance, an ANCOVA controlled by the covariate of age was employed.

Results

Rasch analysis

Summary statistics for the MFRM model can be seen in Table 5. The Wright map is presented in Figure 1.

Table 5
Rasch Summary Statistics

	Test takers	Raters	Items
N	60	2	34
Measures			
Mean	.42	.00	.00
SD (pop.)	1.10	.04	.84
SE	.22	.04	.15
RMSE (pop.)	.23	.04	.15
Adjusted (True) *SD* (pop.)	1.08	.02	.82
Infit *MS*			
Mean	1.08	1.09	1.06
SD (pop.)	.44	.09	0.32
Outfit *MS*			
Mean	1.07	1.09	1.07
SD (pop.)	.44	.07	.33
Homogeneity index (χ 2)	1046.50	2.50	959.80
df	59	1	33
p	.00	.11	.00
Separation (pop.)	4.69	.51	5.41
Reliability of separation (pop.)	.96	.20	.97
Inter-rater reliability			
Observed exact Agreement %		81.0	
Expected %		34.1	
Rasch κ		0.71	

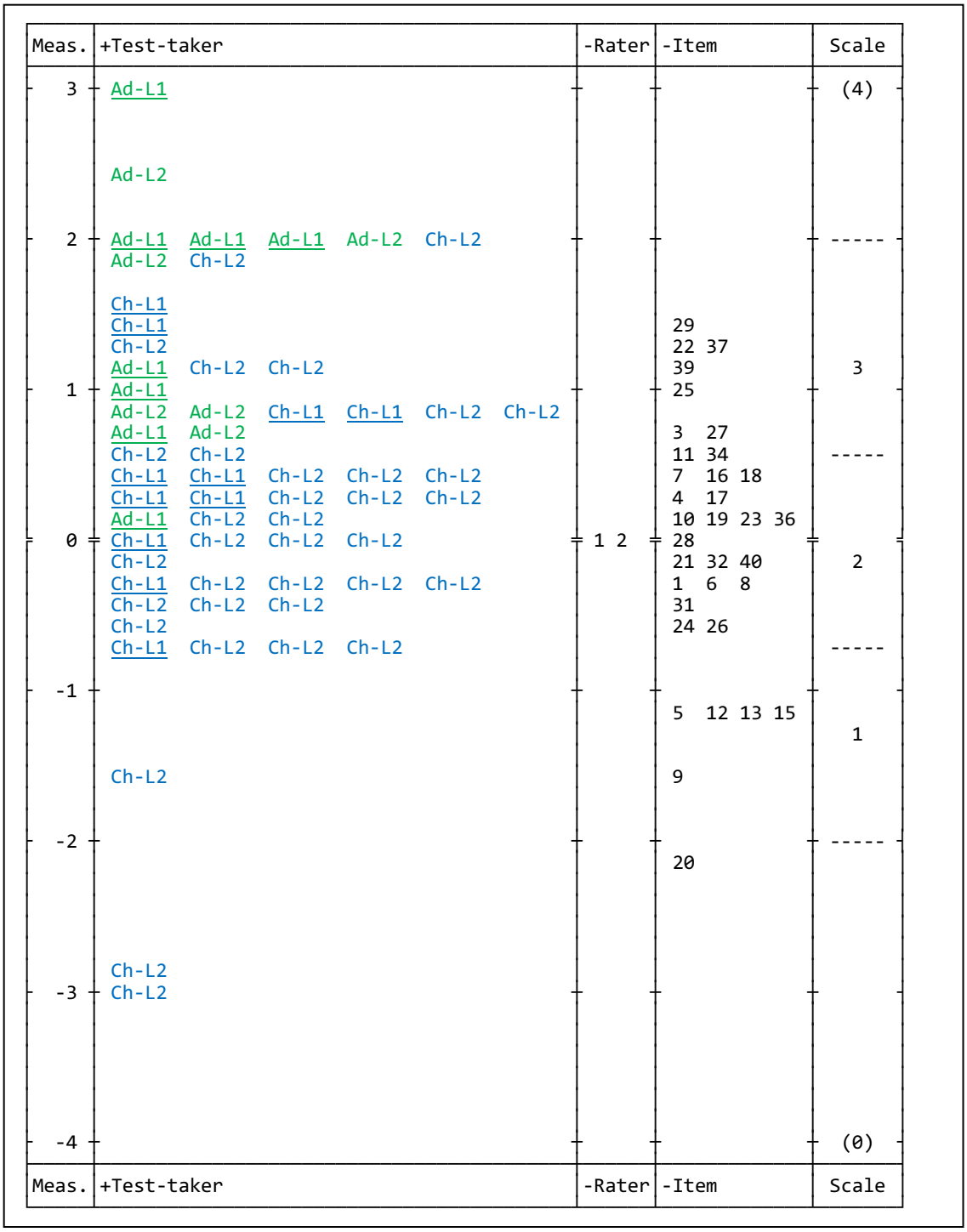

Figure 1. Wright variable map. Adult participants are denoted by "Ad", children by "Ch", and hearing status of parents as "L1" for those with at least one Deaf parent, and "L2" for those with hearing parents. "L1" test-takers are underlined.

As shown, the scoring instrument was able to separate the examinee sample into four distinct levels of ability with a reliability of separation of .96. This can be interpreted similarly to a Cronbach's alpha reliability coefficient (Wright & Masters, 2002), and, as such, the instrument can be understood to be highly reliable, providing evidence of scoring validity (RQ1). Although there were three children and one adult whose abilities were outlying, most

examinees' abilities were grouped around the mean of 0.42 logits, and the distribution of abilities was roughly similar to the distribution of item difficulties.

The raters were very nearly equivalent in severity, with a mean severity of 0 logits and a standard deviation of .04. The reliability of the separation between their severities is .20, and a Chi-square test of their comparative severities is non-significant, indicating that there is virtually no difference between the raters' severities. Finally, the Rasch-*kappa* interrater reliability coefficient of .71 indicates a very high degree of interrater agreement. As such, they can be understood to be rating objectively, and therefore do not present a threat to the scoring validity of the SRT. A pairwise bias analysis revealed that two items (Items 36 and 40) were rated significantly differently by the raters. These items were then subjected to qualitative item analysis to determine if they might represent more complex linguistic structures or were longer than other sentences as a potential explanation for the scoring differences. However, these were not found to be the case, suggesting that the differences were merely spurious (cf. Discussion).

Finally, the items can be separated into five distinct levels of difficulty with a reliability of separation of .97. Average fit statistics are fairly close to their expected values of 1, and the fairly small standard deviations indicate that there was relatively little variation in the degree of fit among the items. After the removal of the six items with Infit MS values over 2.0, the remainder of the items all displayed adequate fit to the Rasch model, demonstrating that they measure the same latent trait, and therefore suggesting construct validity.

Comparative analyses

Comparison to the adult sample

An independent samples *t*-test (Table 6) revealed a significant difference between Child and Adult fair averages, with an effect size in the "large" range, according to the Plonsky and Oswald (2014) thresholds (RQ4). As this difference would be predicted by studies in the field of sign language linguistics (e.g., Rinaldi et al., 2018), this finding lends further support to an argument for context validity.

Table 6
Descriptive Statistics and t-Test for Child and Adult Rasch Fair Averages

	Child	Adult				
N	46	14	*t*	*df*	*p*	*d*
Measures						
Mean	2.16	3.11	-4.61	58	.000	1.56
SD	.72	.47				
SE	.11	.12				

A pairwise bias report (Table 7) revealed seven items which exhibited significantly different difficulty estimates for the Child and Adult samples, with four (Items 11, 16, 36, and 37) being harder for the children and three (Items 18, 22, and 26) being harder for the adults. All effect sizes were "small," according to the Plonsky and Oswald (2014) L2-specific thresholds (see Discussion below).

Table 7
*Pairwise Bias Report for Children and Adults by Item (*p ≤ *.05 only)*

	Child		Adult				Rasch-Welch			
Item	Meas.	*SE*	Meas.	*SE*	Contrast	Joint *SE*	*t*	*df*	*p*	*d*
11	.57	.14	-.48	.39	1.05	.41	2.56	23	.018	.55
16	.47	.14	-.64	.40	1.10	.43	2.59	23	.016	.55
18	.38	.14	1.09	.30	-.70	.34	-2.09	26	.046	.42
22	1.25	.15	1.97	.30	-1.22	.33	-2.15	27	.041	.42
26	-.52	.15	.71	.32	1.31	.35	-3.48	27	.002	.68
36	.11	.15	-1.20	.48	.80	.50	2.63	21	.016	.59
37	1.30	.15	.50	.32	.80	.36	2.23	26	.035	.45

Correlation between the SRT and the Narrative Comprehension Test

With a sub-sample of 19 test-takers, the SRT Rasch fair average scores were correlated with the raw scores of the DSGS narrative comprehension test in order to seek evidence of criterion-related validity (RQ2). The results of the sub-sample, [r = .726, n = 19, p < .001 (2-tailed)], are statistically significant, and the R^2 represents 52.7% shared variance between the two variables. The strength of the correlation can be considered to be strong (>.60; Plonsky & Oswald, 2014).

Chronological age

To address RQ3, a simple linear regression model was calculated to determine the extent to which the fair average scores on the SRT can be predicted by the chronological age of the test-takers (one missing value for age). Age was found to significantly predict SRT performance [F (1, 43) = 17.705, p < .001]. The test-takers' fair average (range: 0.14–3.49) increased in average by .132 for each year of age, and the R^2 indicates that 29.2 % of the variance in the scores is accounted for by age. The effect size (f = .642) benchmark can be considered as a strong effect (f > .40; Cohen, 1988).

Parents' hearing status

To address RQ3 and determine whether Deaf children with at least one Deaf parent performed better on the SRT than Deaf children of hearing parents, a one-way, between-groups analysis of covariance (ANCOVA) was computed, controlled by the covariate chronological age. There was a significant difference in the test performance between the children of Deaf parents and the children of hearing parents [F (1, 42) = 7.27, p = .010,$\eta^2_{partial}$ = .148]. The parental hearing status factor explains only 14.8% of the variance of the fair average. There was also a significant relationship between the covariate chronological age and the dependent variable while controlling for the independent variable with F (1, 42) = 24.14, p < .001, $\eta^2_{partial}$ = .365. Chronological age explains 36.5% of the variance of the test performances of the test-takers (see also Figure 2). The implications of these results will be discussed in the Discussion section.

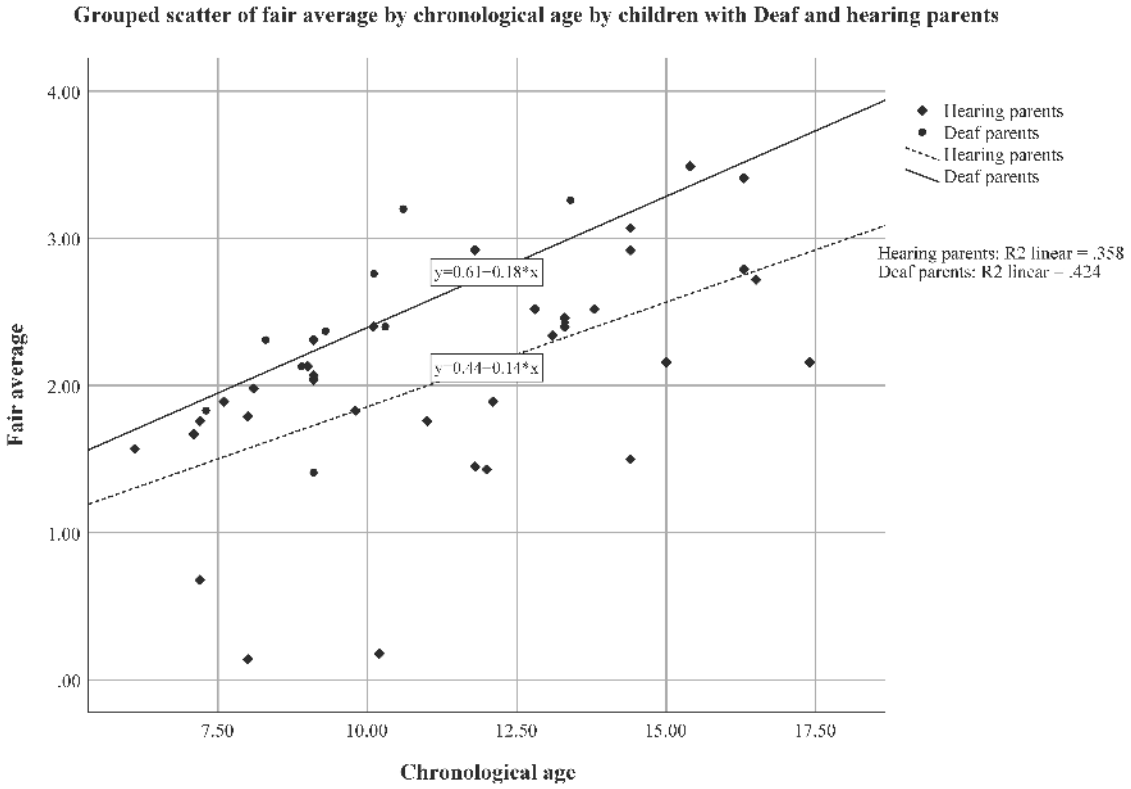

Figure 2. Fair Average of Deaf children with Deaf and Hearing parents, Controlled for Chronological Age

Discussion

The goal of this study was primarily to seek evidence of context, scoring, and criterion-related validity within the socio-cognitive framework for an SRT for DSGS aiming to assess the grammatical development of school-aged Deaf children. Although not empirically demonstrable, the item and rating scale development, and rater training (RQ4) provide some evidence for an argument of context validity, while evidence to support an argument for scoring validity (RQ1) of the SRT was found during the Rasch analysis. It revealed four distinct levels of ability within the sample with a reliability of separation of .96, which can be interpreted similarly to a Cronbach's alpha (Wright & Masters, 2002), indicating very high reliability. Also, five distinct levels of item difficulty with a reliability of separation of .97 were found. The Rasch-*kappa* inter-rater reliability coefficient shows a very high degree of agreement between the two raters.

Two items (Items 36 and 40 of 40 items) were rated significantly differently by the raters. It is striking that both items occur towards the end of the test. A potential source might be (1) fatigue of the raters to explain that they scored these items differently or (2) fatigue by the test-takers making these items harder to score. The first issue (fatigue of raters) has been reported for spoken language assessment too (e.g., Ling, Mollaun, & Xi, 2014). It would have been useful to conduct a follow-up interview with the raters to discuss why they scored these sentences differently (e.g., Isaacs & Thomson, 2013), but the actual scoring took place in summer 2015 and it was therefore not possible to collect any valid follow-up data.

Due to the absence of a test that measures the same construct as the SRT, a sub-sample of the children (n = 19) were also tested on a DSGS narrative comprehension test (Haug & Perrollaz, 2015) in order to seek evidence of criterion-related validity (RQ2). The results of the

correlation can be considered as strong according to the Plonsky and Oswald (2014) threshold. This evidence contributes to a limited argument for the criterion-related validity.

Investigation of the test-taker characteristics' (age, parents hearing status) impact on scores (RQ3) revealed that the test mostly as expected based on the existing literature. The comparison between the performances of the adult users of DSGS ($N = 14$) and the children ($N = 46$) revealed a significant difference, contributing evidence of context validity (RQ4). However, seven items (Items 11, 16, 18, 22, 26, 36, and 37) showed significantly different difficulty estimates for the samples. Of these seven items, four items were more difficult for the children (Items 11, 16, 36 and 37). The authors considered whether the four items that were harder for the children might pose a threat to validity with regard to the test-taker characteristics from an acquisition perspective; i.e., that the items represent specific linguistic structures of DSGS which might not have been acquired by all children and are therefore inappropriate for the intended test-takers. The "problem" with this hypothesis is that, for example, constructed action, which is a discourse strategy using manual and non-manual components to "express a referents actions, utterances, thoughts, feelings and/or attitudes" (Cormier, Smith, & Zwets, 2013) (e.g., Item 11) and which is normally mastered above nine years old (e.g., Morgan, Herman, & Woll, 2002), also occurs in other items (e.g., Items 3, 6, 12). These three items did not differ in difficulty for the two samples. For that reason, although this hypothesis did not bear out in the present study, age should continue to be investigated in future sign language SRT research.

Three of these seven items (Items 18, 22, and 26) were harder for the adults. It is impossible to look at these three items from an acquisition perspective as in the case of the items that were harder for the children (in theory, the adults should outperform the children on all items). Also other potential criteria that might explain performance differences, like complexity or length of the items, cannot really explain the differences between the two samples. Further investigation would be needed in the future to shed some light on the question why these three items were more difficult for the adults than the children.

External variables contributing to the performance differences in the children's sample have been identified in the literature (e.g., Mann, 2006) and set in relation to the SRT results. Chronological age, a crucial variable in child acquisition research, significantly predicted the SRT scores of the children with a strong effect size (Cohen, 1988). This provides further evidence of criterion-related validity, as scores should be expected to increase with age and, therefore, with linguistic development and acquisition.

The variable parents' hearing status (analogous to L1/L2), when controlled for age, predicted score differences between children with Deaf and hearing parents, but explained only 14.8% of the variance in scores, in contrast with previous work on German Sign Language assessment (Haug, 2011). This may or may not contribute to an argument for validity. On the one hand, one would expect those who have grown up using DSGS with their parents to exhibit more facility with it, in which case, this result is somewhat surprising. On the other, however, given the young age of the participants, it may simply be the case that overall linguistic development is a much better predictor of performance on an integrated test task such as an SRT, requiring sufficient experience and facility with the language to not only parse input, but recreate it in response. Clearly, further work with more varied samples is needed.

Conclusion

This study has reported the results of the development and evaluation of an SRT for DSGS for the purpose of demonstrating scoring and criterion-related validity (RQ1 and RQ2), to ensure that test-taker characteristics impacted known factors that explain the performance of the children (RQ3), and also demonstrating context validity (RQ4). Although some issues may

require further examination (e.g., difference in the scoring of the raters on four items; why some items are too difficult for the children), the results demonstrate evidence of context, scoring, and criterion-related validity with regard to global DSGS proficiency and development, and, furthermore, provide a basis for continuing development of the SRT in question, and for encouraging others to consider using SRTs for sign language assessment.

The study does, however, suffer some limitations, chief of which was the relatively small sample size. It would have also been preferable to ensure that all performances were double-rated, although the raters in the present study appeared to operate virtually indistinguishably. Additionally, more background information of the test-takers should be collected, for example, information about the test-takers non-verbal IQ and working memory skills in order to investigate if and to which degree cognitive resources are required to solve the task of an SRT (e.g., for spoken languages: Bartlett, 2018).

Likely future directions include a closer examination of the rating scale, by comparing the results of the 5-step scale of the present study to shorter and dichotomous scales, as these are more frequently found in the literature. The test remains in development, and further validation studies are underway.

Overall, this study provides important validation work in a lesser tested language, Swiss German Sign Language, and offers insight into the use of SRTs for sign language assessment generally. The work can be used as a template for other researchers working in similar contexts to develop and validate their sign language test.

Acknowledgements

We would like to express a special thanks to our colleague Martin Venetz, University of Applied Sciences for Special Education Zurich, one of the co-authors of this paper who passed away on May 9, 2018. Martin provided us with continuous methodological support over the course of the project and contributed greatly to this paper.

Additionally, we also would like to thank our colleagues who shared their sentence repetition tests with us: Peter Hauser and colleagues from the Rochester Institute of Technology, USA; Katherine Rowley, University College London (UCL); Chloe Marshall, UCL Institute of Education, London; Christian Rathmann, Humboldt-Universität zu Berlin; Okan Kubus, University of Applied Sciences Magdeburg-Stendal; Krister Schönström, Stockholm University; and Pasquale Rinaldi and colleagues from the CNR, Rome. We also would like to thank all of our raters and test administrators, the schools for the Deaf, and the parents who were involved in this project.

References

Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3–20. https://doi.org/10.1177/0265532214531254

Batty, A. O., & Haug, T. (forthcoming). *Validating the rating scale for a sentence repetition test for Swiss German Sign Language*. Unpublished manuscript.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

Bartlett, A. N. (2018). *The association between sentence repetition and other cognitive abilities in school-aged children*. Master of Arts dissertation, University of Windsor, Canada. Retrieved from https://scholar.uwindsor.ca/etd/7496

Boyes Braem, P. (1995). *Einführung in die Gebärdensprache und ihre Erforschung* (Vol. 11). Hamburg: Signum-Verlag.

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36,* 3–18. https://doi.org/10.1016/j.asw.2018.02.003

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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. R. (2012, March). *The British Sign Language Sentence Reproduction Test: Exploring age-of-acquisition effects in British deaf adults*. Paper presented at the 34. DGfS Jahrestagung, Frankfurt am Main. Retrieved from http://dgfs.uni-frankfurt.de/dgfs/ag13_07_1800-1830_Cormier.html

Cormier, K., Smith, S., & Zwets, M. (2013). Framing constructed action in British Sign Language narratives. *Journal of Pragmatics*, *55*, 119–139. https://doi.org/10.1016/j.pragma.2013.06.002

Devescovi, A., & Caselli, C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language & Communication Disorders*, *42*(2), 187–208. https://doi.org/10.1080/13682820601030686

Ebling, S. (2016). *Automatic translation from German to synthesized Swiss German Sign Language* (Dissertation). Universität Zürich, Zürich. Retrieved from http://www.cl.uzh.ch/dam/jcr:8c0f6d30-05dc-4e31-9324-0ed7ef74214b/ebling_diss.pdf

Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585–602. https://doi.org/10.1177/0013164408323240

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491. https://doi.org/10.1093/applin/aml001

Erlenkamp, S. (2012). Syntax: Aus Gebärden Sätze bilden. In H. Eichmann, M. Hansen, & J. Hessmann (Eds.), *Handbuch Deutsche Gebärdensprache - Sprachwissenschaftliche und anwendungsbezogene Perspektiven* (pp. 165–198). Hamburg: Signum-Verlag.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*(2)*,* 419–447. https://doi.org/10.1111/lang.12157

Graham, C. R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected proceedings of the 2008 second language research forum* (pp. 57–72). Somerville, MA: Cascadilla Press.

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

Hauser, P., Supalla, T., & Bavelier, D. (2008). American Sign Language sentence reproduction test: Development and implications. In R. Müller de Quadros (Ed.), *Sign Languages: Spinning and unraveling the past, present and future* (pp. 160–172). Florianopolis, Brazil: Editora Arara Azul. Petrópolis.

Haug, T. (2011). Methodological and theoretical issues in the adaptation of sign language tests: An example from the adaptation of a test to German Sign Language. *Language Testing*, *29*(2), 181–201. https://doi.org/10.1177/0265532211421509

Haug, T., & Perrollaz, R. (2015). DSGS-Verständnistest einer Geschichte (DSGS-GV).

Herman, R. (2002). *Assessment of BSL development*. Unpublished doctoral dissertation, City University London.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

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

Johnston, T., & Schembri, A. (2007). *Australian Sign Language: An introduction to sign language linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511607479

Jones, R. (1994). Sentence repetition: A useful oral language screening device. *Australasian Journal of Special Education*, *18*(2), 21–27. https://doi.org/10.1017/S1030011200023174

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

Kubus, O., & Rathmann, C. (2012, March). *Degrees of difficulty in the L2 acquisition of morphology in German Sign Language*. Paper presented at the 34. DGfS-Tagung, Frankfurt am Main. Retrieved from http://dgfs.uni-frankfurt.de/dgfs/ag13_07_8800-8800_Kubus.html

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*(12), 3423–3430. https://doi.org/10.1016/j.ridd.2014.08.026

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago, IL: MESA Press.

Linacre, J. M. (2018). Facets (Version 3.80.4). Retrieved from http://www.winsteps.com/facets.htm

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, *3*(4) 479–499. https://doi.org/10.1177/0265532214530699

Mann, W. (2006). *Examining German deaf children's understanding of referential distinction in written German and German Sign Language (DGS)*. Doctoral dissertation, San Francisco State University & University of California, Berkeley, San Francisco, CA.

Marshall, C., Mason, K., Rowley, K., Herman, R., Atkinson, J., Woll, B., & Morgan, G. (2015). Sentence repetition in deaf children with specific language impairment in British Sign Language. *Language Learning and Development*, *11*(3), 273–251. https://doi.org/10.1080/15475441.2014.917557

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

Meir, N., Walters, J., & Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. *International Journal of Bilingualism*, *20*(4), 421–452. https://doi.org/10.1177/1367006915609240

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/10.1353/sls.2004.0005

Morgan, G., Herman, R., & Woll, B. (2002). The development of complex verb constructions in British Sign Language. *Journal of Child Language*, *29*, 655–675. https://doi.org/10.1017}S0305000902005184

Morgan, G., & Woll, B. (Eds.). (2002). *Directions in sign language acquisition – Trends in language acquisition research*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tilar.2

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, *4*(4), 386–422.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In N. Miyake, D. Peebles, & R. S. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 2132–2137). Austin, TX: Cognitive Science Society.

O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13–32). Basingstoke: Macmillian.

Padden, C. (1990). The relationship between space and grammar in ASL verb morphology. In C. Lucas (Ed.), *SLR theoretical issues* (pp. 118–132). Washington, D.C.: Gallaudet University Press.

Pfau, R., & Quer, J. (2010). Nonmanuals: Their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign languages* (pp. 381–403). New York: Cambridge University Press. https://doi.org/ 10.1017/CBO9780511712203.018

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research: Effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Poll, G. H., Betz, S. K., & Miller, C. A. (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research*, *53*, 414–429. https://doi.org/10.1044/1092-4388(2009/08-0016)

Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics*, *19*(2), 109–127. https://doi.org/10.1080/02699200410001669870

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

Sarandi, H. (2015). Reexamining elicited imitation as a measure of implicit grammatical knowledge and beyond...? *Language Testing, 32*(4), 485–501. https://doi.org/10.1177/0265532214564504

Schönström, K., & Holmström, I. (2017, May). *Elicited imitation tasks (EITs) as a tool for measuring sign language proficiency in L1 and L2 signers*. Paper presented at the 6th International ALTE conference: Learning and assessment: Making the connection, Bologna.

SignMET Consortium (2016). *Deliverable 5.3: Final report that summarizes the results of the study on sign language abilities in children, taking into account the role of specific factors related to deafness.* Unpublished scientific report. EU project SignMET.

Singleton, J. L., & Newport, E. L. (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

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*(3), 723–751. https://doi.org/10.1111/lang.12129

Sutton-Spence, R., & Woll, B. (1999). *The linguistics of British Sign Language: An introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139167048

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Weir, C. (2005). *Language testing and validation: An evidenced-based approach*. New York, NY: Palgrave MacMillan. https://doi.org/10.1057/9780230514577

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions*, *16*(3), 888.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528. https://doi.org/10.1177/0265532215594643