Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 2 (1), 20-40 (2019)
https://dx.doi.org/10.29140/lea.v2n1.85

Sign Language Learning and Assessment in German Switzerland: Exploring the Potential of Vocabulary Size Tests for Swiss German Sign Language

TOBIAS HAUG [ab]
Email: tobias.haug@slas.ch

SARAH EBLING [a]
Email: sarah.ebling@hfh.ch

PENNY BOYES BRAEM [ac]
Email: boyesbraem@gmail.com

KATJA TISSI [a]
Email: katja.tissi@hfh.ch

SANDRA SIDLER-MISEREZ [a]
Email: sandysidler@gmail.com

[a] *Interkantonale Hochschule fuer Heilpaedagogik, SWITZERLAND*
[b] *Sign Language Assessment Services, SWITZERLAND*
[c] *Centre for Sign Language Research Basel, SWITZERLAND*

Abstract

In German Switzerland the learning and assessment of Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) takes place in different contexts, for example, in tertiary education or in continuous education courses. By way of the still ongoing implementation of the Common European Framework of Reference for DSGS, different tests and assessment procedures are currently been developed and their potential is explored to support the learning and assessment of DSGS. Examples of this are two vocabulary size tests. The first is a web-delivered Yes/No Test, the second a Translation Test from written German to DSGS. For both tests, the same set of items was used. The items were sampled from DSGS teaching materials. For the development of the two vocabulary size tests, 20 DSGS adult learners of ages 24 to 55 ($M = 39.3$) were recruited as test takers. An item analysis of the test results yielded candidates for removal from the item set. Cronbach's Alpha showed good results for both tests (>.90), and inter-rater reliability of the Translation Test also indicated promising results (Cohen's Kappa = .613, $p < .001$). Evidence contributing to content validity was collected based on the sampling method of the test items. Due to the lack of a second DSGS vocabulary test that could be used to establish concurrent validity, external variables were identified and investigated as possible external criteria contributing to the performance of the test takers. One variable, number of courses attended, showed a significant correlation with the test results.

Keywords: Swiss German Sign Language (DSGS), sign language assessment, vocabulary size tests

Data Availability Statement: *All relevant data are within this paper.*

Introduction

The implementation and the use of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2009) is a rather new development in the field of learning sign languages as a second or foreign language in tertiary education in Europe. It has only been with recent attempts to align sign language curricula to the CEFR that the development of assessment instruments to evaluate adult learners of a sign language has become possible. Evidence for this are European projects such as *D-Signs* (Leeson & Grehan, 2009) and *ProSign: Sign Language for Professional Purposes* (Leeson, Haug, Rathmann, Van den Bogaerde, & Sheneman, 2018).

We find evidence of progress also in the German-speaking part of Switzerland, where aligning existing Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) curricula to the CEFR has become an important topic and, subsequently, the assessment of adult learners has gained more attention. In Switzerland, three different sign languages are used (Boyes Braem, Haug, & Shores, 2012). In the German-speaking area, DSGS is the primary language of approximately 5,500 Deaf[1] sign language users and a second language to approximately 13,000 hearing persons (Boyes Braem, 2012a). The group of hearing learners includes hearing children of Deaf adults, sign language interpreters, teachers for the Deaf as well as people who use DSGS due to a wide variety of other professional or personal reasons.
There are different DSGS course providers in Switzerland. By far the largest provider is the Swiss Deaf Association, the national umbrella organization for the Deaf, which offers standard courses on seven different levels. Each level corresponds to 30 contact hours that extend over one semester. Currently, the most pressing issue is the lack of reliable and valid DSGS tests that could be used to assess learners in these courses. So far, no DSGS tests for adult learners are available.

DSGS learning takes place in German Switzerland in different contexts, for example, in the sign language interpreter training program in Zurich, family sign language classes, and courses for teachers of the Deaf. Generally speaking, there is a lack of reliable and valid assessment procedures to support DSGS learning. In order to overcome this general lack, different tests and assessment procedures have been developed in two exploratory studies. On the one hand, the Sign Language Proficiency Interview (Caccamise & Samar, 2009) has been adapted to DSGS (Haug, Nussbaumer, & Stocker, 2019), on the other, two vocabulary size tests have been developed for DSGS. The development and evaluation of these two vocabulary tests are the focus of this paper.

One aspect rendering development and research of sign language tests difficult is that most sign languages are under-documented and under-resourced and therefore have, e.g., no corpora or reference grammars at their disposal. This is also true for DSGS: No balanced and representative DSGS corpus exists, and no reference grammar has been put forth. Research on sign language test development and use is itself a rather young field within the domains of sign linguistics and applied linguistics (Haug, 2015).

Literature Review: Sign Language Linguistics and Vocabulary Assessment

Sign Language Structure

Sign Language Phonology

The articulators in sign languages have been divided into two distinct categories: *manual* and *non-*

[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).

manual components (e.g., Boyes Braem, 1995). The manual components are the hands and the arms. Non-manual components include several features of the face (mouth, cheeks, eyes, nose, eyebrows, eye gaze) as well as positions and movements of the head and the upper torso (Boyes Braem, 1995; Sutton-Spence & Woll, 1999). Manual and non-manual components are usually produced simultaneously.

The manual components are traditionally considered the smallest building blocks of *sign language phonology*. They were first investigated in American Sign Language (ASL) by Stokoe (1960) and later extended by Battison (1978) and Klima and Bellugi (1979). These sub-lexical manual units, which together compose a sign, are the handshape, location, movement, and hand orientation.

Non-manual components that are produced with the mouth can refer to the lip movements of a spoken word, termed *mouthings*, and have been found to be common in most European sign languages. One of their uses is to distinguish between signs that have the same manual form (manual homonyms), as with the DSGS signs BRUDER (‘brother’), SCHWESTER (‘sister’), and GLEICH (‘same’). Another non-manual component, eye gaze, can be used, e.g., to re-establish reference in signing space, and raised eyebrows can be added to manual signs, e.g., to distinguish an interrogative sentence from a declarative sentence (Pfau & Quer, 2010).

Sign Language Lexicon

Johnston and Schembri (2007) proposed a model for the organization of the mental lexicon in sign languages. They divide the mental lexicon into a *native* and a *non-native* part. The native lexicon is further divided into a *conventional* and a *productive lexicon* (Figure 1). The conventional lexicon consists of signs (lexical types) that have a stable form-meaning relationship, e.g., the German Sign Language (*Deutsche Gebärdensprache*, DGS) sign AUTO (‘car’), which can be used in different contexts without a change in meaning (König, Konrad, & Langer, 2012).

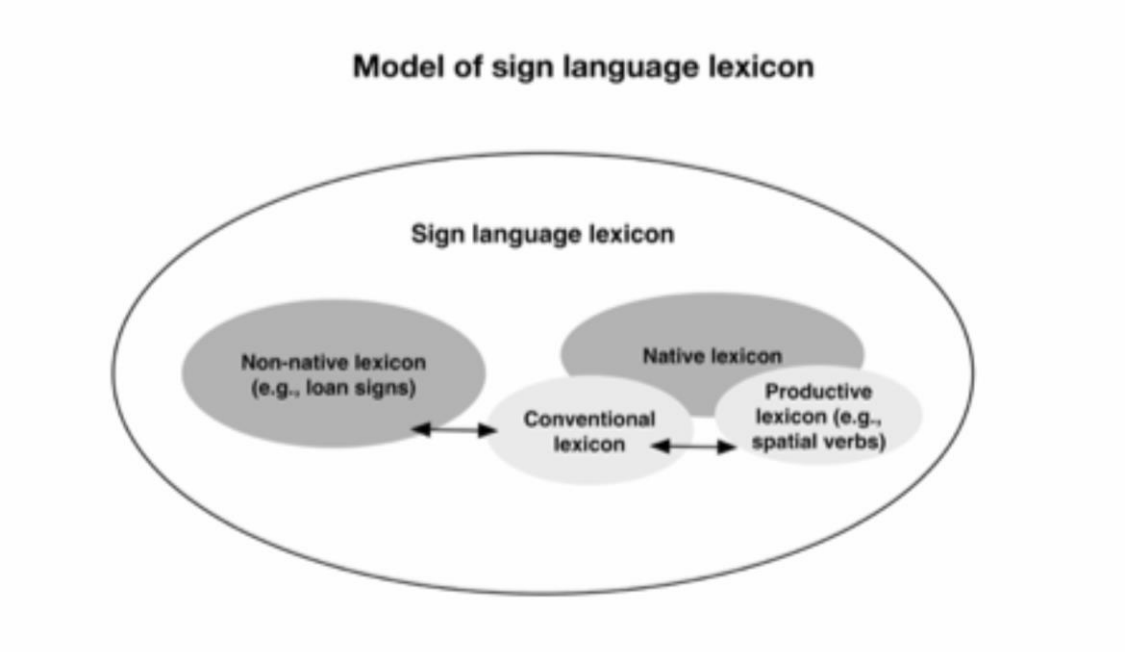

Figure 1 Model of sign language lexicon (Johnston & Schembri, 2007)

The productive lexicon is quite different and does not consist of an easy-to-determine number of signs.

Sign forms labeled as productive are realized and understood in a given context to convey a specific meaning. The signs themselves are not conventionalized, but their sub-lexical units (especially the handshapes) are. A slight change in one sub-lexical unit can change the meaning. Because of the multiple possibilities for varying the parameters of the form, no citation entry/base form of signs in the productive lexicon is possible.

The number of conventional sign types of a sign language is difficult to determine: Estimates range from 2,500 to 5,000 signs for Australian Sign Language (Auslan) and DGS, respectively (Johnston & Schembri, 1999; Ebbinghaus & Heßmann, 2000). Since there are a potentially large number of context-specific meanings, the size of the productive lexicon cannot be determined.

For the purpose of the present research, only signs of the native, conventional lexicon were considered for the vocabulary tests. In order to arrive at a concept comparable to that of *word families* (Read, 2000), signs that involve morphological changes to the lexical base form could also be included (Johnston & Schembri, 2007). However, the problem remains that this group of signs is less clearly defined for sign languages than for spoken languages such as English, which would have an impact on the definition of what a correctly produced sign in a DSGS vocabulary test is. Just considering sign types that are known to have stable form-meaning relationships (König *et al*., 2012) is further complicated by the fact that little research exists on acceptable phonetic variations of signs (see, e.g., Arendsen, 2009). For DSGS, a recent study investigated acceptable variants of lexical signs in L1 and L2 users of DSGS (Ebling *et al*., 2018), which also served as the basis for defining the criterion of correctness for the translation test of interest in this article.

From a linguistic perspective, the visual-spatial modality of sign languages has an impact on the test design and scoring criteria when the entire lexicon - productive and conventional signs - would be considered as the construct of a vocabulary test for DSGS. In order to arrive at a comparable construct of a vocabulary size test as for English (e.g., Read, 2000), only the conventional lexicon was considered for the current study.

Aspects of Vocabulary Knowledge: Size and Depth

A distinction often applied to vocabulary knowledge is that between *vocabulary size* and *vocabulary depth* (e.g., Read, 2000; Schmitt, 2014). While size is comparatively straightforward to define, namely as the number of words (or word families) a learner knows, for example, in English (e.g., Meara, 1996; Read, 2000; Schmitt, 2014), and is often tested through vocabulary size tests, vocabulary depth is defined as how well someone knows a word. The concept of vocabulary depth is commonly divided into different parts (e.g., Read, 2000). A well-known approach, following Nation (2001), is to distinguish between form, meaning, and use, with features of orthography, phonology, morphology, syntax, collocations, and pragmatics, with respect to both receptive and productive skills.

The distinction between vocabulary size and depth can also be applied to DSGS, but with some modifications due to the absence of a widely accepted conventionalized writing system for sign languages (Boyes Braem, 2012b). For the purpose of this project, the focus will be on the development of a vocabulary size measure.

Different Forms of Vocabulary Assessment Instruments

Frequently used instruments for receptive and productive vocabulary assessments are checklists (e.g., Yes/No tests), matching tests (e.g., receptive items in which a test taker is asked to match a target word with other related words or short definitions), or translation from the L1 into the L2 or vice versa (Read, 2007; Kremmel & Schmitt, 2016). For the purpose of the project at hand, a productive vocabulary test (L1L2 Translation Test) was included in addition to a Yes/No Test (YN Test). A YN Test format was preferred over a matching test because the latter format would have required the production of a larger number of video materials in DSGS (target signs and a number of matching answers) and a more complex technical implementation which was constrained by the project budget.

Yes/No Tests

YN Tests have been reported as a widely used measure for receptive vocabulary size (e.g., Read, 1993; Beeckmans, Eyckmans, Janssens, Dufranne, & Van der Velde, 2001; Pellicer-Sánchez & Schmitt, 2012; Stubbe & Stewart, 2012; Beglar & Nation, 2014). They have often been used as placement tests (e.g., Laufer & Nation, 1999; Read, 2000; Laufer & Goldstein, 2004; Harrington & Carey, 2009;) or diagnostic tests (e.g., Read, 2007; Sevigny & Ramonda, 2013). These kinds of tests are time-effective, easy to administer, and allow for testing a large number of words in a short amount of time (Read, 1993, 2007; Meara, 1996; Harrington & Carey, 2009; Stubbe & Stewart, 2012). Completing a YN Test places only limited demands on the test taker (Pellicer-Sánchez & Schmitt, 2012). The basic design of YN Tests consists of a test taker seeing a word and indicating whether he or she knows the word (e.g., Beeckmans *et al*., 2001). Items of YN Tests can be sampled from word frequency lists, like the Academic Word List (e.g., Gardner & Davies, 2013), or different 1,000-level word lists from the British National Corpus (e.g., Nation, 2004).

YN Tests have been subject to criticism in the past. One point of criticism is the possibility of overestimation (i.e., "falsely claiming knowledge of real words", Stubbe *et al*., 2010, p. 4) and guessing on the part of the test takers, i.e., a test taker can rate more words with "yes" than he or she actually knows (e.g., Read, 1993, 2000). In order to compensate for this effect, Anderson and Freebody (1983) added pseudowords to their YN Test. Pseudowords are phonologically possible forms of a language (Mochida & Harrington, 2006). An example for a pseudoword is *borth* (based on the English word *birth*; Mochida & Harrington, 2006, p. 82) The terms *pseudowords* and *nonwords* are sometimes used interchangeably (for a discussion see Sevigny & Ramonda, 2013). For the DSGS YN Test in the present study, the term *nonsense signs* was used (Mann, Marshall, Mason, & Morgan, 2010). Nonsense signs are phonologically plausible forms of a sign language that bear no meaning in that language. Nonsense signs can be created by using signs from other sign languages, where the combination of the used sub-lexical units compose a possible form in DSGS, but do not exist in DSGS and bear no meaning. For example, for the present study, the Italian sign for the concept 'play' is a possible phonological form that could exist in DSGS but bears no meaning.

Scores of YN Tests have often been employed in conjunction with other measures of vocabulary size knowledge (e.g., Vocabulary Levels Test; Nation, 2001) to control for the responses of the YN Test (e.g., Stubbe *et al*., 2010; Pellicer-Sánchez & Schmitt, 2012; Stubbe, 2012, 2015). The picture that emerges is very diverse. For example, Meara and Buxton (1987) report a statistically significant correlation (r = .703) between the scores of a YN Test and a vocabulary multiple-choice test. A similar result is reported by Anderson and Freebody (1983) (r = .84) and Mochida and Harrington (2006), who applied different scoring methods for the YN Test with a range of significant correlations with the Vocabulary Levels Test (r = .85 to .88). However, other studies show a different picture. For example, Cameron (2002) reports no significant correlation between YN Test scores and scores from the Vocabulary Levels Test across different 1,000-word levels, with Spearman rho ranging from .15 to .45. Eyckmans (2004) observed similar results in that no significant correlations between YN Test scores and translations test scores were found (r = .03 to .05).

In a YN Test, the words a test taker chooses as "known" are called *hits* (e.g., a test taker indicates that he or she knows the English word 'car'), while words rated as "unknown" are termed *misses* (e.g., a test taker indicates that he or she does not know the English word 'family'). Pseudowords rated as known are considered *false alarms* (e.g., a test taker indicates that he or she knows the pseudoword 'borth'), and pseudowords checked as unknown are *correct rejections* (e.g., a test taker indicates that he or she does not know the pseudoword 'borth'; Mochida & Harrington, 2006; Stubbe, 2015).

As Meara (2005) states, "[t]he real difficulty […] is not the production of the tests, but how we interpret the scores that the tests generate" (p. 278). The simplest solution is to combine the hits and correct rejections into a total score (Mochida & Harrington, 2006). However, according to Huibregtse, Admiraal, and Meara (2002), this approach is problematic, as the hit rate is an indicator of vocabulary knowledge, while the correct rejections (or false alarms) are an indicator for the amount of guessing. Three different approaches have been proposed for dealing with the false alarm rate (Schmitt, 2010; Stubbe, 2015):

(1) Setting a maximum amount of false alarm responses as a threshold for exclusion of a test taker. For example, Schmitt, Jiang, and Grabe (2011) suggested a 10% acceptance rate (three out of 30 items), and Stubbe (2012) applied a 12.5% rate.
(2) Adjusting the YN scores by using one of several different proposed formulas (see Huibregtse *et al*., 2002 and Stubbe, 2015). The simplest formula is to subtract the false alarm rate from the hit rate to arrive at a score that better reflects vocabulary knowledge (*true score*).
(3) Applying a regression model to use the YN test scores to predict scores of translation tests (Stubbe & Stewart, 2012).

No consensus exists as to what is the best approach or which formula in Approach 2 works best (Schmitt, 2010).

Regarding the false alarm rate across different studies, the range is considerable. For example, Mochida and Harrington (2006) report less than 5% false alarms in a study with English L2 university students; Stubbe (2012) reports a false alarm rate of a little over 4% in Japanese learners of English at university level; Harrington and Carey (2009) report a false alarm rate of 17% in a study of English L2 learners in Australia, and Eyckmans (2004) reports 13-25% false alarms in French learners of Dutch at different levels.

Beeckmans *et al*. (2001) note that there are no clear guidelines regarding the ratio of real words to pseudowords. For example, among the ratios reported in different studies are 30:3 (Schmitt & Zimmermann, 2002), 90:60 (Mochida & Harrington, 2006), 96:32 (Stubbe, 2015), 60:40 (Meara & Buxton, 1987; Eyckmans, 2004), 72:28 (Harrington & Carey, 2009), and 40:20 (Meara, 1992).

Translation Tests and Scoring Issues

Productive tests have been used as a means for verifying the test takers' self-reported vocabulary in a YN Test (e.g., Stubbe *et al*., 2010; Stubbe, 2015). The study at hand consisted of developing and applying a productive test. The simple form of a Translation Test is that an L1 word is provided and the test taker produces the L2 translation (Laufer & Goldstein, 2004). In the DSGS vocabulary test used for the present study, the L2 translation consisted of a DSGS sign which will be produced by the test taker.

The responses were scored manually by two raters. Stewart (2012) cautions that production tests that are hand-scored can result in an inconsistency between raters. An important issue concerning the development of scoring instruments is (1) to define a criterion of correctness and (2) to decide whether two (i.e., right/wrong) or more degrees of correctness (i.e., partial credit) should be applied (Bachman & Palmer, 1996). For assessing a single area of language knowledge, such as vocabulary, the right/wrong distinction can be useful, but when different areas of language knowledge are assessed, partial credit might be appropriate (Bachman & Palmer, 1996).

Research Questions

Based on the goal of this study, which was to develop and evaluate two vocabulary size tests for DSGS, the following main research questions were investigated:

(1) Do the two vocabulary tests show evidence of reliability?
(2) Do the two vocabulary tests show evidence of validity?

Methodology

Development of Instruments

Four instruments were developed:

(1) Yes/No Vocabulary Test for DSGS (YN Test);
(2) L1/L2 Vocabulary Translation Test for DSGS ("Translation Test");
(3) Scoring instrument for the Translation Test; and
(4) A background questionnaire for test takers.

The visual-spatial modality of sign languages has an effect on the test design. Due to the absence of a conventional written form for sign languages including for DSGS (Boyes Braem, 2012b), stimuli of the test items need to be videos (YN Test) and the responses need to be recorded with a video camera (Translation Test).

Item Selection

Due to the absence of a large corpus for DSGS, it was not possible to create a corpus-based frequency list of DSGS signs as it exists for English (e.g., Laufer, Elder, Hill, & Congdon, 2004) that could be used as the basis of a vocabulary test. Instead, a list of 98 DSGS vocabulary items (including 5 practice items) was used that had been developed as part of a Swiss National Science Foundation project (Ebling *et al*., 2018). The aim of this ongoing project is to develop an automatic sign language recognition system to employ within a DSGS vocabulary test for the CEFR level A1. The items used in the test were sampled from existing DSGS teaching materials (Boyes Braem, 2004a, 2004b, 2005a, 2005b) known to correspond to level A1. The number of sign types available in the DSGS teaching materials is approximately 3,800 (Boyes Braem, 2001). To reduce this number to 98 items, the following linguistic criteria were applied (Ebling *et al*., 2018):

(1) Remove name signs, i.e., signs for persons (e.g., CHARLIE CHAPLIN), organizations (e.g., name of a university), and places (e.g., country names), as many of these are borrowed from other sign languages.
(2) Remove body-part signs like NASE ('nose'), as these are often produced by merely pointing at the respective body part, i.e., using an "indexing technique."
(3) Remove pronouns like DU ('you'), as they also correspond to indexical signs.
(4) Remove number signs, as they tend to have several regional variants, e.g., the number sign ELF ('eleven').
(5) Remove signs making use of fingerspelling, like the sign JANUAR ('January'), which involves the letter J from the DSGS manual alphabet.
(6) Remove signs composed of multiple successive elements, as most of these signs also occurred in the DSGS teaching materials as separate lexemes. For example, the sign ABENDESSEN ('dinner') is composed of the two signs ABEND ('evening') and ESSEN ('meal'), both of which are also contained in the list of sign types of the DSGS teaching materials.
(7) Remove old signs, as current DSGS learners cannot be expected to know them.

(8) Remove productive forms. The reason for this step was that the phonological parameters of productive signs tend to be variable, which poses a great challenge to the sign recognition system that is part of the assessment framework in the Swiss National Science Foundation project mentioned above.
(9) Remove signs appearing in fewer than four of the five DSGS dialects.
(10) Reduce manual homonymy: Since the goal was to have as many different sign forms in the vocabulary test as possible, form-identical signs were identified (e.g., BRUDER ('brother'), SCHWESTER ('sister'), and GLEICH ('same')) and only one chosen for the test. Preference was given to that sign which was contained in a list of 1,000 common sign concepts (Efthimiou *et al*., 2009).
(11) Remove signs that are very similar to well-known co-speech gestures, such as the sign SUPER, which corresponds to a thumb-up gesture.
(12) Remove signs with German glosses that are lexically ambiguous. For example, the German word AUFNEHMEN can have the meaning of *record* or *accept/include*, concepts which in DSGS are expressed with two separate signs. In cases like these, test takers confronted with the German gloss AUFNEHMEN would not know which sign to produce.
(13) From the resulting pool of signs, concepts that also occurred in studies investigating familiarity or subjective frequency ratings for BSL (Vinson, Cormier, Denmark, Schembri, & Vigliocco, 2008) and ASL (Mayberry, Hall, & Zvaigzne, 2013; Caselli, Sehyr, Cohen-Goldberg, & Emmorey, 2017) and in the list of 1,000 sign concepts (Efthimiou *et al*., 2009) were prioritized. In this way, the 3,800 sign types from the DSGS teaching materials were reduced to a set of 98 test items.

The item set was not balanced with respect to parts of speech, as is often done when sampling items for a spoken language vocabulary test. This was because the question of whether the concept of parts of speech can be applied to sign languages is a highly debated one within sign linguistics (see, e.g., Erlenkamp, 2001).

Yes/No Test for DSGS

The same items were used for the YN Test and the Translation Test. As verification of test takers' knowledge is not possible (Stubbe, 2015), 25 nonsense signs were developed to control for the self-report aspect of the YN Test. This resulted in a total of 123 items (5 practice items + 93 test items + 25 nonsense signs). All signs were video-recorded for the YN Test. Following the five practice items placed at the beginning, the order of the remaining signs was randomized as in the original list from the project.

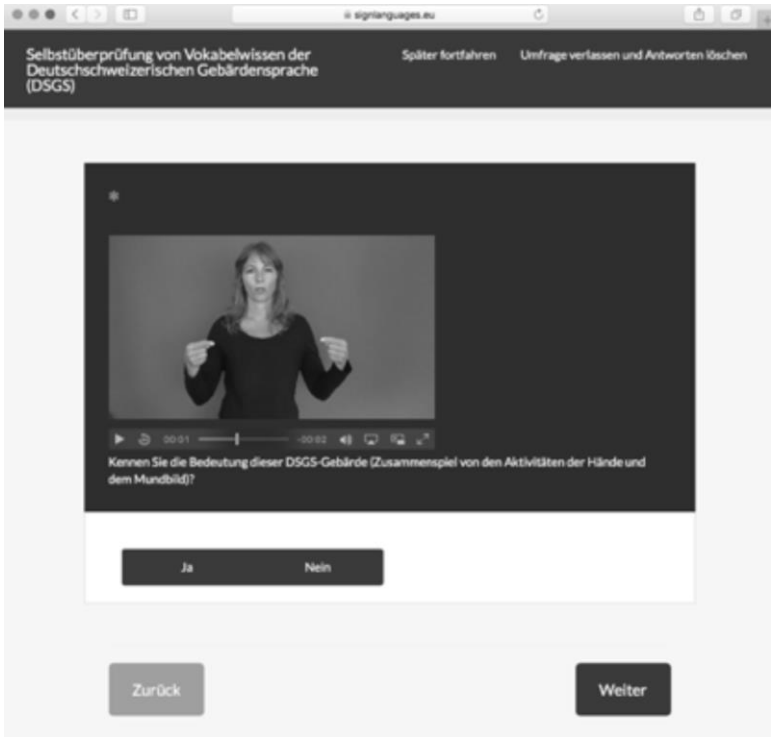

Figure 2 Yes/No Test Practice Item 1 WIDERSPRUCH ('contradiction')

For the YN Test, a Limesurvey installation was used. Limesurvey (http://www.limesurvey.org) is an open-source software for conducting surveys. Each sign was presented separately with the question "Do you know the form and the meaning of the sign?," with a "Ja"/"Nein" ('yes'/'no') button underneath, a "Weiter" ('continue') button to proceed to the next item (Figure 2), and a "Zurück" ('back') button to return to the preceding item. All data were automatically sent to a secure server for storage. The test was self-administered, but the researcher was present during test taking in case there were any technical issues.

L1/L2 Translation Test and Scoring Instrument for DSGS

For the Translation Test, the same set of concepts was used, delivered in the same order as in the YN Test but omitting the nonsense signs, which resulted in 98 signs. The test was delivered as a PowerPoint presentation on a laptop. Following an instruction in written German on the first slide, on each slide, the German word appeared, and a German sentence designed to disambiguate the meaning was provided. For example, to avoid confusion of German homonyms like "Schloss," which can have the meanings of *castle* and *lock* that are expressed by two different signs in DSGS, the following example sentence was provided: "In Lenzburg steht ein Schloss." ('There is a castle in Lenzburg [name of a Swiss city].'). The test taker was seated at a table facing a video camera and with the laptop placed on the table to the right- or left-hand side of the test taker, depending on his or her preference. A test administrator clicked through the slides so that the test taker could look directly at the video camera while producing the sign.

The criterion of correctness (Bachman & Palmer, 1996) was defined as the accuracy of the translation of the German word into DSGS. Building on the work of Ebling *et al*. (2018) on acceptable variants of lexical DSGS signs, the resulting categories informed the criterion of correctness. For a correctly produced DSGS sign, a score of "1" was assigned, for an incorrect form, a score of "0." Apart from the decision of whether the sign was correct or not, an additional category "no sign produced" was

introduced. Test takers also received a score of "0" for these null productions, but the data were collected separately from the incorrect signs to allow for a more fine-grained analysis of inter-rater reliability. The video-recorded data were scored independently by two raters.

Background Questionnaire

The background questionnaire included questions in the following areas: (1) General background information (e.g., name, gender, age); (2) language and language-learning background, including variables such as first and preferred language(s), DSGS courses attended, and self-judgment of receptive and productive DSGS skills; (3) background information on professional/vocational training, work context (i.e., current position, use of DSGS at work), and test takers' use of DSGS in their free time; and (4), Deaf family members and, if present, language of communication. It took an average of ten minutes to complete the questionnaire.

Testing Procedure

After a pilot study with three test takers, which resulted in some small revisions to the background questionnaire and the wording of the YN Test instructions, a main study was conducted with 20 test takers. The test takers were sampled through different professional networks of the researchers. The testing protocol consisted of the following phases: (1) conveying background information on the project, (2) filling in online background questionnaire (test takers), (3) completing the YN Test (online), and (4) completing the Translation Test (this included video-recording the test takers for later analysis).

Sample of Main Study

Of the 20 test takers, five were male and 15 female. The test takers were between 24 and 55 (M = 39.3) years old at the time of testing. 19 of the 20 test takers were hearing; one wore a cochlear implant and had acquired German as her first language but was also learning DSGS as an adult. The majority of the test takers had one spoken language (e.g., a Swiss German dialect or Standard German; n = 18) as their L1. Two participants reported having grown up with two spoken languages. All participants had learned DSGS as adults (age of acquisition range: 18-53 years, M = 35.4).

Background of the Raters

Both raters are female. Rater 1 (R1) is 54 years old, Rater 2 (R2) 45. Both raters acquired DSGS from birth through Deaf family members and use DSGS in both their private and their work context. Both are trained sign language instructors and have been teaching and evaluating different groups of learners (beginning learners as well as sign language interpreting students) for over 30 (R1) and nearly 20 years (R2), respectively. R1 is a lecturer in the sign language interpreting program at the University of Applied Sciences for Special Needs Education (HfH) and has also been involved in many different DSGS research projects over the last 20 years. R2 works also as research assistant at the HfH and for the Swiss Deaf Association. Both raters received specific training to use the scoring instrument of the Translation Test.

Statistical Procedures

The data collected for this project comprise:

(1) Test scores of the YN Test
(2) Test scores of the Translation Test
(3) Questionnaire data

Statistical Assumptions

The data of the YN Test and the Translation Tests were normally distributed, allowing for the use of parametric procedures. A statistical significance level of .05 (2-tailed) was used. For the YN Test, the adjusted scores (*true scores*) served as the basis for all of the analyses. For analyses that involved the YN Test score, one study participant was excluded, as she/he had to terminate the YN Test after the first 40 items due to technical difficulties. The scores of the Translation Tests of Rater 1 and Rater 2 were treated separately.

Nonparametric statistics were chosen for investigating inter-rater reliability of the Translation Test and to compute correlations that involved ordinal-level data. In addition, Cohen's (1992) scale for determining the strength of the relationship of a correlation was chosen: (1) .10 to .29 as small, (2) .30 to .49 as medium, and (3) .50 to 1.0 as large. Furthermore, the coefficient of determination, i.e., how much shared variance the two variables have in common, was calculated by squaring the r or r_s value (Pallant, 2016).

The analysis of effect size for a paired-samples t-test followed Cohen's (1988) scale of .01 as a small effect, .06 as a moderate effect, and .14 as a large effect by applying the eta squared formula.

Statistical Procedures

For research question 1 an (1) item analysis, (2) Cronbach's Alpha, and (3) Cohen's Kappa were calculated. For research question 2, correlations (Pearson, Spearman rank) were calculated.

Results

Item Analysis of the Test Scores

Items were retained in the item pool when they met the following criteria: (1) facility value (*FV*) between .20 to .90 (e.g., Bachman, 2004) and/or (2) a corrected item-total correlation (*CITC*) >.30 (Carr, 2011; Green, 2013).

Item Analysis of the YN Test Acores

A total of 26/98 items did not meet the defined criteria of *FV* and/or *CITC* and were thus candidates for future removal.

Item Analysis of the Translation Test Scores

As for Rater 1, 16 items were candidates to be removed from the test, and 24 items would need to be removed based on the results of Rater 2. Across both raters, 26 items were identified that did not meet the *FV* and/or *CITC* criterion. As the Translation Test is still under development, it was decided that only items for which at least one of the criteria for removal defined above was met by both raters were considered to be discarded. This resulted in a total of twelve items (Table 1).

Table 1 *Items to be Removed from the Translation Test and YN Test*

Items (DSGS gloss)	Translation	*FV* R1	*CITC* R1	*FV* R2	*CITC* R2	*FV* YN Test	*CITC* YN Test
P1 WIDERSPRUCH	'contradiction'	.15	.281	.15	.274	N/A	N/A

I7 SCHÜTZEN	'to protect'	.2	.266	.15	.12	N/A	N/A
I14 TELEFONIEREN*	'to call'	1.0	.0	.55	.277	1.0	.0
I25 SAMMELN	'to collect'	.55	.26	.45	.194	N/A	N/A
I27 BLAU*	'blue'	.9	.242	.75	.159	1	0
I28 FREUND*	'friend'	.9	.122	.8	-.095	.94	.254
I33 EI*	'egg'	.65	.05	.5	-.079	.56	-.004
I36 MONAT	'month'	.6	.22	.6	.207	N/A	N/A
I43 VERKAUF	'sale'	.35	.246	.15	-.143	N/A	N/A
I52 FRAGEN*	'to ask'	.85	.171	.85	.185	.94	.116
I61 FARBE*	'color'	.8	.13	.8	.171	.94	-.203
I84 ABEND*	'evening'	1.0	.0	.9	-.028	1.0	.0

*Items where at least one of the two criteria occurred across the Translation Test scores (Rater 1 and Rater 2) and the YN Test scores.

Evidence for Reliability

Cronbach's Alpha

For the YN Test, the Cronbach's Alpha value for the real DSGS signs as shown in Table 2 can be considered high (α = .980), and for the nonsense signs, acceptable. The Cronbach's Alpha value for the translation test can also be considered high, with a small difference between Rater 1 (α = .970) and Rater 2 (α = .961).

Table 2 *Cronbach's Alpha in the YN Test and the Translation Test Scores*

Test	Cronbach's Alpha
YN Test: Real signs, 98 items (*n* = 18*)	.980
YN Test: Nonsense signs, 25 items (*N* = 19)	.760
Translation Test, Rater 1, 98 items (*N* = 20)	.970
Translation Test, Rater 2, 98 items (*N* = 17*)	.961

*Listwise deletion based on all variables in the procedure

Inter-Rater Reliability of the Translation Test

Cohen's Kappa was applied to investigate inter-rater reliability (Gwet, 2014). Instances of the "no signs produced" category were treated as missing values in order to gain a more realistic picture of the raters' behavior (in an operational test, occurrences of "no sign produced" would be treated as "wrong").

The agreement between the two raters on the translation test was κ = .613, p <.001, which is considered "substantial" according to Landis' and Koch's (1977) Kappa benchmark scale. To investigate the raters' strictness, a paired-samples t-test was applied. Using the raw scores as the basis for the comparison, there was a statistically significant difference between the ratings of Rater 1 (M = 53.6, SD = 22.96) and Rater 2 (M = 47.7, SD = 20.73), t (19) = 5.871, p <.001 (2-tailed). The mean difference was 5.9, with a 95% confidence interval ranging from 3.79 to 8.0. Effect size was calculated using the eta squared formula (Pallant, 2016). This resulted in a value of .644, which qualifies as a "strong" effect according to Cohen (1988) indicating a large magnitude of the difference between the raw scores of Rater 1 and Rater 2. More precisely, Rater 1 evaluated the test takers less strictly than did Rater 2.

Evidence Contributing to the Vocabulary Tests' Validity

Based on the information obtained from the test takers' background questionnaire, the variable "number of DSGS courses attended" proved to be statistically relevant. More precisely, the results yielded a statistically significant relationship between the number of DSGS courses attended and the YN Test scores ($n = 17$; $r_s = .528$, $n = 17$, $p = .036$, (2-tailed)). The amount of shared variance was .278 or 28% between the two variables, which was not high. The strength of the correlation can be considered as strong (>.50).

The results between the Translation Test scores and the variable "courses attended" were equally statistically significant (Table 3). This means that the more courses the test takers had taken, the higher their scores were. The correlation was slightly higher with Rater 2 than with Rater 1. Both correlations can be considered strong (>.50). The shared variance of the two variables with Rater 1 was .337 or 33.7%, and with Rater 2, .361 or 36.1%.

Table 3 *Correlation Translation Test Scores and Number of Courses Attended, by Rater (n = 17)*

Rater	Spearman's rho (r_s)	*p*
Rater 1	.581	.014*
Rater 2	.601	.011*

*statistically significant at .05 level, 2-tailed.

Comparison of the Performances of the Test Takers on both Vocabulary Tests

In order to investigate evidence for a statistical relationship between the scores of the YN Test and the Translation Test, two correlations were calculated. Both correlations were strong (>.50), i.e., the higher test takers scored on one of the tests, the higher they scored on the other test. The shared variance of the two variables was .657 or 65.7% for Rater 1 and .752 or 75.2% for Rater 2 (Table 4).

Table 4 *Correlation YN Test Scores and Translation Test Scores, by Rater (N = 19)*

Rater	Pearson's *r* (*r*)	*p*
Rater 1	.811	.001*
Rater 2	.867	.001*

*statistically significant at .001 level, 2-tailed.

Nonsense Signs of the YN Test: False Alarm Rate

The frequency distribution of the nonsense signs was analyzed. Seven test takers claimed to not know any of the nonsense signs (correct rejections), the remaining test takers claimed to know between one and ten nonsense signs (false alarms).

The hit rates of the real items ($N = 98$) and the nonsense signs ($N = 25$) are displayed in Table 5. The false alarm rate was 9.26% (ranging from 0 to 10 signs).

Table 5 *Hits and False Alarms of the YN Test (N = 19)*

	Hits (max. Score = 98)	False Alarms (max. Score = 25)
M Scores	64.82	2.32
Rate (in %)	66.12%	9.26%
Range of Scores	17-98	0-10
SD	24.15	2.65

Discussion

Item Analysis of the YN Test and Translation Test

Given the developmental stage of both tests and the fact that they had to have the same set of items, it was decided to remove only items that met at least one of the two criteria of *FV* and *CITC* in both the translation and the YN test scores. This resulted in a total number of seven items, marked with an asterisk in Table 1. For future operational versions of both tests, the order of the items will be changed according to the results of the level of difficulty (balanced between Rater 1 and 2 in the translation task and checked with the *FV* of the YN test scores), starting with easier items.

Inter-Rater Reliability: Cohen's Kappa

The inter-rater reliability score for Rater 1 and Rater 2 as reported in the Result section is acceptable for the purpose of this study. However, it is desirable to have a higher agreement value in future studies. The results of the paired-samples t-test showed that Rater 1 scored less strictly than Rater 2 on average. It was not possible to determine the precise sources of the disagreement (and also not possible to include a third rater), but possible explanations could be (1) the lack of a more intense rater training (e.g., McNamara, 1996; Fulcher, 2014) and/or (2) insufficient clarity as to the concept of correctness of a sign, which also touches on the ongoing discussion of what acceptable variants are (Ebling *et al*., 2018). These issues are currently investigated in a follow-up study whose aim is to apply a Rasch analysis to the data and qualitative interviews with the raters (Haug, Batty, & Ebling, 2019).

At first glance, a "correct" or "incorrect" decision at the single-sign level does not seem to pose a challenge, as judging the correctness of an entire signed utterance would. However, as mentioned earlier, there is to date no consensus as to what an acceptable phonetic variant of a sign is. This might be influenced by the fact that sign languages are not standardized languages (Adam, 2015). The previously mentioned study of Ebling and colleagues (2018) on acceptable variants in DSGS will help to shed more light on this topic. This will in turn also help to get a clearer definition of the proposed criteria of correctness (Bachman & Palmer, 1996).

In the future, one would also need to decide whether a combined test score should be reported if a certain minimum of agreement between two or more raters has been reached, and how the scores will be communicated to the test takers. Intra-rater reliability should also be investigated in the future.

Evidence of Validity in the Two Tests

Content Validity

For the study at hand, content validity was established through the way the items were sampled. The list of lexical DSGS signs, compiled within the previously mentioned project, was sampled from well-established DSGS teaching materials (Boyes Braem, 2004a, 2004b, 2005a, 2005b) that are used for beginning learners in different DSGS learning contexts in German Switzerland, and to which specific linguistic criteria were applied to reduce the number of items from around 3,800 to 98 (Ebling *et al*., 2018). This provided a solid basis for developing the two tests here. Even though the aspects of frequency and part of speech could not be accounted for due to the lack of a corpus and sufficient research for DSGS, respectively, the list used represents signs that beginning learners are exposed to in their DSGS classes. However, the absence of a DSGS corpus poses a big problem in the long term, when the items will eventually need to be modified or replaced. A future study, therefore, consists of letting a larger group of experienced Deaf sign language instructors rate the levels of difficulty of a larger set of items. This will allow future sampling of items from different levels of difficulty.

Test Performances on the Two Tests

External Variables Contributing to the Test Takers' Performance

As pointed out in the introduction, due to the lack of an existing DSGS vocabulary test for adult learners that would allow for investigating concurrent validity, external variables (self-assessment of DSGS skills, number of DSGS courses attended, and DSGS learning contexts) were identified through the background questionnaire that might explain or contribute to the overall test scores. Of these three variables, only the variable "number of courses attended" exhibited a statistically significant correlation with the scores of both tests. A possible explanation for the fact that the other two variables did not correlate with the test scores is that a self-assessment is too difficult for the test takers to undertake, due to the lack of a reference. For example, "very good" DSGS skills might mean something different to a beginning learner of DSGS than to a trained sign language interpreter. The self-assessment rating might also be influenced by the level of self-criticism of the learner. As for the learning context variable, it is one that is hard to quantify, i.e., what exactly does using DSGS with friends or at work imply in terms of language exposure? The number of courses is perhaps the most appropriate variable to correlate with the test scores in that it is easy to quantify and also contains the notion of language exposure.

In addition, the applied statistics were only of correlational nature, which means it is uncertain whether the number of courses is really the predictor variable contributing to the test scores. Nevertheless, the correlation between the number of courses attended and the test scores represents evidence that an external variable can be used as argument for concurrent validity. This method of approaching concurrent validity has previously been applied in sign language testing by Mann (2006) and Haug (2011).

Comparing Test Scores of Both Tests

The correlation of the test scores of the Translation Test and the YN Test as reported in Table 4 showed promising results with a strong correlation. These results are comparable to studies on English that correlated YN test scores with scores of existing vocabulary tests, both with correlations (e.g., Anderson & Freebody, 1983; Mochida & Harrington, 2006) and by computing the shared variance (Mochida & Harrington, 2006). However, there are also studies for spoken language that report weaker correlations (e.g., Cameron, 2002; Eyckmans, 2004).

Evaluation of the Nonsense Signs

The false-alarm rate was somewhere in-between what has been reported for spoken language YN tests in the literature. In the future, it will be interesting to investigate whether the false alarm rate is a function of over- or underestimation with respect to the translation scores of the test takers.

The literature for spoken language YN tests provided a framework for developing the YN Test for DSGS, containing (1) the concept of pseudowords (e.g., Beeckmans *et al*., 2001), (2) suggestions for adjusting the YN Test raw scores by means of a correction formula (e.g., Stubbe, 2015), and (3) a reference regarding the false alarm rate (e.g., Stubbe, 2012). However, there is one crucial difference between nonsense signs and pseudowords: In DSGS, the form and the meaning of a sign is composed of both manual and non-manual components that are produced at the same time, i.e., two different visual channels are used simultaneously to produce a linguistic symbol. This use of simultaneity on different linguistic levels in sign languages is quite different from how spoken languages are structured, especially at the level of the isolated lexical item. Test takers in the pilot study reported that they applied different strategies in cases where they did not understand a sign as a whole (i.e., information

from all channels). It was only when the manual form was unknown to them that they tried to retrieve meaning from the mouthing. Because of this, it is clear from this study that more research is needed to determine whether the nonsense signs really fulfill their function for a YN test. One could pursue this question by letting L1 users of DSGS judge whether nonsense signs could potentially be DSGS signs under the following conditions: (1) with no mouthing, (2) accompanied by a German mouthing, and (3) accompanied by a mouthing from the original sign language (e.g., the BSL sign KNOW and the English mouthing /know/). For future YN Test uses, one could also experiment with including the information in the instructions that there are some signs in the test that are not actual DSGS signs.

Conclusion

The goal of the study reported in this article was to develop and evaluate two DSGS vocabulary size tests, one relying on a self-report format (YN Test) and one on a verifiable format (Translation Test), with both tests showing the potential of being used operationally in different DSGS learning contexts as placement and/or diagnostic instruments for beginning adult learners in German Switzerland (along with other DSGS instruments). Developing and evaluating a test of vocabulary knowledge for a language like DSGS that is under-documented and under-resourced poses a number of methodological challenges, which have been discussed in this article. Despite these constraints, it was possible to develop and evaluate the two tests.

A limitation of the work reported in this study was the sample size of the main study, which at 20 was too small to allow for generalization of the findings. A third rater would have been preferable in order to obtain more data, and the rater training was rather short. In addition, the effectiveness of the nonsense signs could not be assessed exhaustively.

This is the first and only study so far in the field of sign language assessment that explicitly addresses vocabulary assessment for adult learners of sign language not only in Switzerland but internationally. Studies from the field of spoken language assessment provided a framework for the development of some aspects of the two tests and were supplemented by studies from sign language linguistics. The results of this study complement studies from the larger field of spoken language assessment and will contribute to future research in sign language testing and assessment and its application.

References

Adam, R. (2015). Standardization of sign languages. *Sign Language Studies*, *15* (4), 432-445. https://doi.org/10.1353/sls.2015.0015

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hudson (Ed.), *Advances in reading/language research: A research annual* (pp. 231-256). Greenwich, CT: JAI Press.

Arendsen, J. (2009). *Seeing signs: On the appearance of manual movements in gestures*. Technische Universiteit Delft, Delft, The Netherlands.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Battison, R. (1978). *Lexical borrowing in American Sign Language*. Silver Spring, MD: Linstok Press.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van der Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18* (3), 235-274. https://doi.org/10.1177/026553220101800301

Beglar, D., & Nation, P. (2013). Assessing vocabulary. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 172-184). Hoboken, NJ: John Wiley & Sons, Inc.

Boyes Braem, P. (1995). *Einführung in die Gebärdensprache und ihre Erforschung [Introduction into sign language research]* (Vol. 11). Hamburg: Signum.

Boyes Braem, P. (2001). A multimedia bilingual database for the lexicon of Swiss German Sign Language. *Sign Language & Linguistics*, *4* (1/2), 133-143. https://doi.org/10.1075/sll.4.12.10boy

Boyes Braem, P. (2004a). *Gebärdensprachkurs Deutschschweiz, Stufe 1. Linguistischer Kommentar [Swiss German Sign Language, level 1: Linguistic commentary]*. Zürich: GS-Media/Schweizerischer Gehörlosenbund SGB.

Boyes Braem, P. (2004b). *Gebärdensprachkurs Deutschschweiz, Stufe 2. Linguistischer Kommenta [Swiss German Sign Language, level 2: Linguistic commentary]r*. Zürich: GS-Media/Schweizerischer Gehörlosenbund SGB.

Boyes Braem, P. (2005a). *Gebärdensprachkurs Deutschschweiz, Stufe 3. Linguistischer Kommentar [Swiss German Sign Language, level 3: Linguistic commentary]*. Zürich: GS-Media/Schweizerischer Gehörlosenbund SGB.

Boyes Braem, P. (2005b). *Gebärdensprachkurs Deutschschweiz, Stufe 4. Linguistischer Kommentar [Swiss German Sign Language, level 4: Linguistic commentary]*. Zürich: GS-Media/Schweizerischer Gehörlosenbund SGB.

Boyes Braem, P. (2012a). *Overview of research on the signed languages of the deaf*. Lecture, University of Basel.

Boyes Braem, P. (2012b). Evolving methods for written representations of signed languages of the deaf. In A. Ender, A. Leemann, & B. Waelchli (Eds.), *Methods in contemporary linguistics* (pp. 411-438). Berlin: De Gruyter Mouton.

Boyes Braem, P., Haug, T., & Shores, P. (2012). Gebärdenspracharbeit in der Schweiz: Rückblick und Ausblick [Sign language research and application in Switzerland: Review and outlook]. *Das Zeichen*, *90*, 58-74.

Bochner, J. H., Samar, V. J., Hauser, P. C., Garrison, W. M., Searls, J. M., & Sanders, C. A. (2016). Validity of the American Sign Language Discrimination Test. *Language Testing*, *33* (4), 473-495. https://doi.org/10.1177/0265532215590849

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6* (2), 145-173. https://doi.org/10.1191/1362168802lr103oa

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, *49* (2), 784-801. https://doi.org/10.3758/s13428-016-0742-0

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155-159.

Council of Europe (2009). *Common European Framework of Reference for languages: Llearning, teaching, assessment*. Cambridge; Strasbourg: Cambridge University Press; Council of Europe.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford; New York: Oxford University Press.

Ebbinghaus, H., & Heßmann, J. (2000). Leben im Kommunikationskonflikt: Zur Ungleichsprachigkeit Hörender und Gehörloser [Different ways of communicating of deaf and hearing people]. In E. Hess-Lüttich & H. W. Schmitz (Eds.), *Botschaften verstehen: Kommunikationstheorie und Zeichenpraxis. Festschrift für Helmut Richter* (pp. 47-66). Frankfurt am Main: Peter Lang.

Ebling, S., Camgöz, N. C., Boyes Braem, P., Tissi, K., Sidler-Miserez, S., Hatfield, … Magimai-Doss, M. (2018). SMILE Swiss German Sign Language data set. *11th Language resources and evaluation conference (LREC 2018)*, 4221-4229.

Efthimiou, E., Fotinea, S. E., Vogler, C., Hanke, T., Glauert, J., Bowden, R., … Segouat, J. (2009). Sign language recognition, generation, and modelling: A research effort with applications in deaf communication. In C. Stephanidis (Ed.), *Universal access in human-computer interaction.*

Addressing diversity, lecture notes in computer science (Vol. 5614, pp. 21-30). Berlin: Springer.

Erlenkamp, S. (2001). Lexikalische Klassen und syntaktische Kategorien in der Deutschen Gebärdensprache: Warum das Vorhandensein von Verben nicht unbedingt Nomen erfordert [Lexical classes and syntactic categories in German Sign Language: Why the existence of verbs does not require nouns]. In H. Leuninger & K. Wempe (Eds.), *Gebärdensprachlinguistik 2000 – Theorie und Anwendung: Vorträge vom Symposium Gebärdensprachforschung im Deutschsprachigen Raum, Frankfurt a.M., 11.-13. Juni 1999 [Sign language linguistics 2000 – Theory and Application]*. (Vol. 37, pp. 67-91). Hamburg: Signum.

Eyckmans, J. (2004). *Measuring receptive vocabulary size: Reliability and validity of the yes/no vocabulary test for French-speaking learners of Dutch*. LOT, Utrecht.

Fulcher, G. (2014). *Testing second language speaking*. London: Routledge.

Gardner, D., & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics*, *35* (3), 305-327. https://doi.org/10.1093/applin/amt015

Green, R. (2013). *Statistical analyses for language testing*. Basingstoke, UK: Palgrave Macmillan.

Gwet, K. L. (2014). *Handbook of inter-rater reliability*. Gaithersburg, MD: Advanced Analytics.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37* (4), 614-626. https://doi.org/10.1016/j.system.2009.09.006

Haug, T. (2011). *Adaptation and evaluation of a German Sign Language test - A computer-based receptive skills test for deaf children ages 4-8 years old*. Hamburg: Hamburg University Press. Retrieved from http://hup.sub.uni-hamburg.de/purl/HamburgUP_Haug_Adaption

Haug, T. (2015). Use of information and communication technologies in sign language test development: Results of an international survey. *Deafness & Education International*, *17* (1), 33-48. https://doi.org/10.1179/1557069X14Y.0000000041

Haug, T., Batty, A., & Ebling. S. (2019). *Investigating raters' behavior on a Swiss German Sign Language vocabulary test using Rasch analysis*. Manuscript in preparation.

Haug, T., Nussbaumer, D., & Stocker, H. (2019). Die Bedeutung der Kognition beim Gebärdensprachdolmetschen [The role of cognition in sign language interpreting]. *Das Zeichen*, *111*, 130-143.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19* (3), 227-245. https://doi.org/10.1191/0265532202lt229oa

Johnston, T., & Schembri, A. (1999). On defining lexeme in a signed language. *Sign Language & Linguistics*, *2* (2), 115-185. https://doi.org/10.1075/sll.2.2.03joh

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

Klima, E., & Bellugi, U. (1979). *Signs of language*. Cambridge, MA: Harvard University Press.

König, S., Konrad, R., & Langer, G. (2012). Lexikon: Der Wortschatz der DGS [The lexicon in German Sign Language]. In H. Eichmann, M. Hansen, & J. Heßmann (Eds.), *Handbuch Deutsche Gebärdensprache: Sprachwissenschaftliche und anwendungsbezogene Perspektiven [Handbook of German Sign Language: Linguistic and applied persepctives]* (Vol. 50, pp. 111-164). Seedorf: Signum.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13* (4), 377-392. https://doi.org/10.1080/15434303.2016.1237516

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorial data. *Biometrics*, *33*, 159-174.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21* (2), 202-226. https://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54* (3), 399-436.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16* (1), 33-51. https://doi.org/10.1177/026553229901600103

Leeson, L., & Grehan, C. (2009). A Common European Framework for sign language curricula? D-Sign(ing) a curriculum aligned to the Common European Framework of Reference. In M. Mertzani (Ed.), *Sign language teaching and learning - Papers from the 1st symposium in applied sign linguistics* (Vol. 1, pp. 21-33). Bristol: Centre for Deaf Studies, University of Bristol.

Leeson, L., Haug, T., Rathmann, C., Van den Bogaerde, B., & Sheneman, N. (2018). *Survey report from the ECML project ProSign: Sign languages for professional purposes. The implementation of the Common European Framework of Reference (CEFR) for signed languages in higher education - Results of an international survey*. https://doi.org/10.13140/RG.2.2.26818.07365

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

Mann, W., Marshall, C. R., Mason, K., & Morgan, G. (2010). The acquisition of sign language: The impact of phonetic complexity on phonology. *Language Learning and Development*, *6* (1), 60-86. https://doi.org/10.1080/15475440903245951

Mayberry, R. I., Hall, M. L., & Zvaigzne, M. (2013). Subjective frequency ratings for 432 ASL signs. *Behavior Research Methods*. https://doi.org/10.3758/s13428-013-0370-x

McNamara, T. F. (1996). *Measuring second language performance*. London; New York: Longman.

Meara, P. (1992). *EFL vocabulary test*. Swansea, UK: Centre for Applied Language Studies.

Meara, P. (1996). The dimension of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. S. Butler, M. de los Á. Gómez González, & S. M. Doval-Suárez (Eds.), *The dynamic of language use* (Vol. 140, pp. 271-285). Amsterdam: John Benjamins.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142-154. https://doi.org/10.1177/026553228700400202

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23* (1), 73-98. https://doi.org/10.1191/0265532206lt321oa

Morgan, G., & Woll, B. (Eds.). (2002). *Directions in sign language acquisition - Trends in language acquisition research*. Amsterdam: John Benjamins.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge; New York: Cambridge University Press.

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). Amsterdam: John Benjamins.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Berkshire, UK: Open University Press.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes-No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29* (4), 489-509. https://doi.org/10.1177/0265532212438053

Pfau, R., & Quer, J. (2010). Nonmanuals: Their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign languages* (pp. 381-403). Cambridge: Cambridge University Press.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10* (3), 355-371. https://doi.org/10.1177/026553229301000308

Read, J. A. S. (2000). *Assessing vocabulary*. Cambridge; New York: Cambridge University Press.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7* (2), 105-125.

Schmitt, N. (2010). *Researching vocabulary*. London: Palgrave Macmillan. Retrieved from http://link.springer.com/10.1057/9780230293977

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64* (4), 913-951. https://doi.org/10.1111/lang.12077

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95* (1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Zimmermann, C. B. (2002). Derivate word forms: What do learners know? *TESOL Quarterly*, *36* (2), 145-171.

Sevigny, P., & Ramonda, K. (2013). Vocabulary: What should we test? In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings* (pp. 701-711). Tokyo: JALT.

Stokoe, W. C. (1960). *Studies in linguistics - Sign language structure: An outline of the visual communication systems of the American deaf*. Buffalo, NY: University of Buffalo.

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in yes/no vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29* (4), 471-488. https://doi.org/10.1177/0265532211433033

Stubbe, R. (2015). Replacing translation tests with Yes/No tests. *Vocabulary Learning and Instruction*, *4* (2), 38-48. https://doi.org/10.7820/vli.v04.2.stubbe

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary tests with linear models. *Shiken Research Bulletin*, *16* (2), 2-7.

Stubbe, R., Stewart, J., & Pritchard, T. (2010). Examining the effects of pseudowords in yes/no vocabulary tests for low level learners. *Language Education and Research Center Journal*, *5*, 1-16.

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

Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., & Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods*, *40* (4), 1079-1087. https://doi.org/10.3758/BRM.40.4.1079

Author Biodata

Prof. Dr. Tobias Haug studied sign linguistics at Hamburg University and Deaf Education at Boston University, where he received his master's in 1998. In 2009, he earned his Ph.D. in sign languages from Hamburg University. In 2017 he completed a distance master in Language Testing, Lancaster University. Since 2004, he has been the program director of and lecturer in the sign language interpreter program at the University of Applied Sciences of Special Needs Education in Zurich.

Dr. Sarah Ebling is a lecturer and researcher at the University of Zurich and University of Applied Sciences of Special Needs Education Zurich. As a computational linguist, her focus is on the contribution of language technology to accessibility. For her Ph.D., she worked on automatic sign language translation and animation.

Penny Boyes Braem (PhD, Dr. h.c.) has conducted studies of L1 acquisition of ASL, the use of "mouthings," characteristics of signed prosody, comparisons of early and later learner signing as well as developing a large databank of Swiss German (DSGS) signs, which is now incorporated into a growing corpus lexicon. (www.fzgresearch.org)

Katja Tissi is a trained sign language teacher for Swiss German Sign Language (DSGS). She works for more than 30 years as a lecturer in the sign language interpreting program at the University of Applied Sciences of Special Needs Education in Zurich. She has been involved in various research projects on DSGS, among others dealing with the creation of technical terms in DSGS, phonological variants, and the assessment in L2 learners.

Sandra Sidler-Miserez is a trained sign language teacher for DSGS. She has been involved in various research project at the University of Applied Sciences of Special Needs Education in Zurich. She is also in charge of the online lexicon for DSGS, maintained by the Swiss Deaf Association.