Language Assessment Quarterly

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hlaq20

Automated Sign Language Vocabulary Assessment: Comparing Human and Machine Ratings and Studying Learner Perceptions

Franz Holzknecht, Sandrine Tornay, Alessia Battisti, Aaron Olaf Batty, Katja Tissi, Tobias Haug & Sarah Ebling

To cite this article: Franz Holzknecht, Sandrine Tornay, Alessia Battisti, Aaron Olaf Batty, Katja Tissi, Tobias Haug & Sarah Ebling (26 Jun 2024): Automated Sign Language Vocabulary Assessment: Comparing Human and Machine Ratings and Studying Learner Perceptions, Language Assessment Quarterly, DOI: 10.1080/15434303.2024.2364877

To link to this article: https://doi.org/10.1080/15434303.2024.2364877

Published online: 26 Jun 2024.

Submit your article to this journal

View related articles

View Crossmark data

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2024.2364877

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS

Check for updates

Automated Sign Language Vocabulary Assessment: Comparing Human and Machine Ratings and Studying Learner Perceptions

Franz Holzknecht[a], Sandrine Tornay[b], Alessia Battisti[c], Aaron Olaf Batty[d], Katja Tissi[a], Tobias Haug[a], and Sarah Ebling[c]

[a]University of Teacher Education in Special Needs (HfH), Zürich, Switzerland; [b]Idiap Research Institute, Martigny, Switzerland; [c]University of Zürich, Zürich, Switzerland; [d]Keio University, Fujisawa, Japan

ABSTRACT

Although automated spoken language assessment is rapidly growing, such systems have not been widely developed for signed languages. This study provides validity evidence for an automated web application that was developed to assess and give feedback on handshape and hand movement of L2 learners' Swiss German Sign Language signs. The study shows good machine-internal and human-machine agreement through many-facet Rasch analysis. Learner perceptions examined through questionnaire responses indicate that the automated system occasionally generated ratings which impacted the quality of feedback at the level of individual signs for individual learners. Implications are discussed from a learning-oriented assessment perspective.

INTRODUCTION

For spoken languages, automated assessment systems have become increasingly common in recent years, for testing both writing and speaking. The main advantage of automated scoring and feedback systems is their cost-efficiency, as large volumes of written and spoken performances can be evaluated in a short amount of time (Van Moere & Downey, 2016). Another advantage of automated scores over human scores is that machines can more reliably rate specific mechanical aspects of language use, such as word frequency, punctuation, or grammatical errors (Davis & Papageorgiou, 2021).

For sign languages, however, automated assessment systems for language production have so far not been developed on a wider scale, despite increased levels of interest by practitioners in this issue (Haug et al., 2023). This has mainly to do with the complexity of collecting and annotating large numbers of signed learner performances, which is a crucial prerequisite as hundreds of rated and annotated performances are usually required to develop an automated scoring system for a specific task (Van Moere & Downey, 2016). While many learner corpora for writing and speaking exist for spoken languages and new task-specific corpora can be created relatively easily (at least for written language), the development of corpora for sign languages is more complex (Fenlon & Hochgesang, 2022). This is because sign

CONTACT Franz Holzknecht franz.holzknecht@hfh.ch University of Teacher Education in Special Needs (HfH), Institute for Language and Communication, Schaffhauserstrasse 239, Postfach 5850, CH-8050 Zürich, Switzerland

languages have no written form and repositories of sign language performances do not exist to the same extent as for written or spoken texts in spoken languages. As sign languages are processed visually and contain both manual and non-manual components, sign language performances first need to be video-recorded before they can be rated and error-annotated by human raters (Hanke & Fenlon, 2022). In addition, video recordings need to be of high enough quality, ideally involving 3-D cameras positioned at multiple angles, to allow for subsequent automated sign language recognition (Ebling et al., 2018). Although signed corpora are predicted to grow in the coming years (Schembri & Cormier, 2022), these practical constraints and associated costs also mean that sign language corpora are generally much smaller than corpora of spoken languages.

Due to these challenges, most existing automated assessment and feedback systems for sign language production are not based on adataset of signed language performances, but on direct comparisons of a learner's video with an L1 signer's reference video. Several such systems have been developed over the last years. For example, Ellis et al. (2016) and Phan et al. (2018) developed a feedback system for Auslan (Australian Sign Language) that rates individual signs recorded through a Microsoft Kinect camera as either correct or incorrect by comparing them to a reference video. The system also provides visual feedback on incorrectly performed signs using avatars. Another system that requires users to be recorded through a Microsoft Kinect camera was developed by Chai et al. (2017). In this application, learners of Chinese Sign Language are assessed on their production of individual signs. They receive an overall score and separate scores for hand movement and handshape, in relation to a reference video. Finally, Liu et al. (2024) describe the development of a sign language recognition and feedback system which only requires a standard camera. Users first learn how to produce a sign and are then evaluated on their production, including feedback on hand position and handshape.

These developments are promising, but certain limitations remain. First and foremost, as outlined above, these systems are not based on a dataset of signed learner performances, but rather on a direct comparison of the learner's production video with one or several reference videos by a single L1 signer. The systems therefore do not account for natural variation between different signers (see also Bayley et al., 2015). For this reason, it is important that sign language recognition and assessment algorithms are based on a representative dataset including both multiple L1 signers and L2 learners, as is the case for automated assessment systems of writing and speaking in spoken languages. Another limitation of some current applications concerns the camera required to record learner performances. The systems which are based on Microsoft Kinect cameras (Chai et al., 2017; Ellis et al., 2016; Phan et al., 2018) need to be updated to other hardware in the future, as the Microsoft Kinect has since been discontinued. More user-friendly solutions requiring less sophisticated camera technology for recording sign language performances, e.g. standard cameras as used by Liu et al. (2017), are necessary to implement automated sign language assessment systems on a larger scale. It should also be noted that all existing systems, including the one described in the current paper, so far only rate the performance on manual components of individual signs. The automated recognition and assessment of continuous signing and of non-manual components such as mouthing, eye-brow movements, head movements, or eye-gaze, which are important parts of sign language communication, is an ongoing research problem, which an interdisciplinary research project is currently working on (see Haug & Holzknecht, 2020).

Apart from these technical and logistical challenges, another aspect which has so far not been addressed sufficiently in the field of automated sign language assessment is the comparison of the systems with human raters. In spoken languages, it is standard to compare automated ratings with a large-enough number of human expert ratings to statistically analyze their agreement and the extent to which they measure the same underlying trait (Van Moere & Downey, 2016). Although limited studies involving only one human rater have been conducted for some of the systems presented above (Chai et al., 2017; Liu et al., 2024), comprehensive studies are still lacking.

In addition, as the main objective of all existing automated sign language assessments is to promote and foster learning, they should also be studied from a learning-oriented assessment (LOA) perspective. LOA has traditionally been under-represented in the assessment literature, and only in the last decade have researchers begun to more systematically focus on the perspectives and insights of the learners themselves (Jin, 2023). Carless and colleagues propose three key components that are central to LOA approaches:

1. *Designing LOA tasks*. The tasks learners need to complete should be authentic and elicit competencies that mirror those needed in real-life.
2. *Developing learners' evaluative expertise*. Learners should develop skills that enable them to critically analyze their work and the work of their peers with a view to support learning.
3. *Learners' engagement with feedback*. Learners need to understand and actively engage with the feedback provided to them so they can use it appropriately to improve learning outcomes.

(Carless, 2015; Carless et al., 2006, pp.964–965).

Automated sign language assessment and feedback systems should ideally also encapsulate these three components. However, it has not been studied whether and to what extent these principles are met in existing systems.

The current paper addresses some of these challenges. Our aim was to validate an automated sign language assessment and feedback system for Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS), which is based on a large dataset of signed L1 and L2 productions, through (1) comparisons of automated ratings with those of human expert raters, and (2) learner feedback questionnaire data on the usefulness of the system with regards to improved learning outcomes. We first briefly describe the system in the next section before outlining the study in detail.

THE WEB SMILE DEMO

The Web SMILE Demo is an application designed for learning and automatically assessing the production of individual lexemes in sign languages. The current version of the Demo was developed in the course of the SMILE (Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment) research projects on DSGS (Ebling et al., 2018; Haug & Holzknecht, 2020). Like the systems outlined above, the Web SMILE Demo currently only provides scores and feedback on manual components of individual lexemes and does not (yet) consider non-manuals. However, in contrast to other existing systems, it is based on a dataset of performances consisting of multiple productions of 100 DSGS lexemes by 9 L1 DSGS signers and 19 L2 DSGS learners, which have been annotated using six category/acceptability criteria at lexeme-level: "(1) same lexeme as target sign (same meaning, same form), (2) same lexeme as target sign (same meaning, slightly different form), (3) same lexeme as target sign (same meaning, different form), (4) same lexeme as target sign (slightly different meaning, slightly different form), (5) different lexeme than target sign (same meaning, different form), (6) different lexeme than target sign (different meaning, different form)" (see Ebling et al., 2018). In addition, the Demo does not require special external camera equipment or specific server capacities but can be run online on any modern computer with an integrated or connected webcam. The Demo currently includes 16 DSGS lexemes, but the number of lexemes will be extended to 100 during the next project (for details about the 16 lexemes, including pictures and links to videos, please see Appendix 1). Although it is only available for DSGS, the Demo's underlying algorithms can be run on data from any sign language. We used the first version of the Demo for the current study (see Tornay et al., 2020).

To practice or be assessed on the production of lexemes with the Web SMILE Demo, learners need a personal computer with a webcam and internet access. After logging in via a web browser, learners first need to sign a data protection consent form. They are then directed to a"How it works" section, which describes the different processes and subpages and includes contact details in case of technical or other problems.

Once they have familiarised themselves with the instructions in the "How it works" section, learners first choose which lexeme they would like to practice. Learners are prompted to produce each lexeme by means of the German language equivalent in written form. Before recording their performance, they can watch a reference video of the target lexeme performed by an L1 user of DSGS (this option can be deactivated for assessment purposes). The current capability of the Web SMILE Demo is thus consistent with the discrete, context-independent approach of conventional form-recall vocabulary tests. After recording their performance, the learner's video is sent to a server for scoring. Processing of feedback and scores currently takes about 10 to 12 seconds (depending on the internet speed), but the aim is to further reduce processing time during the next phase of the project.

After processing, the Demo provides automated feedback on five different aspects of the learner's performance: handshape is assessed as either correct or incorrect for both dominant and non-dominant hand (signers usually have a dominant hand which they use to produce one-handed signs, see Watkins & Thompson, 2017), hand movement is rated on a scale for both hands, and overall score is rated according to the same scale as hand

Table 1. Rating Scale for Hand Movement and Overall Score.

Score	Description
4	little or no divergence from the target sign
3	small divergence from the target sign
2	significant divergence from the target sign
1	large divergence from the target sign
0*	the sign does not correspond to the target sign

*This is only reported for overall score. The degrees of divergence are determined through algorithms based on the category/acceptability annotations of the underlying dataset (see Ebling et al., 2018).

movement. For the overall score, the algorithm uses the handshape and hand movement features of both hands to compute the alignment between the learner's production and the model. The rating scale for hand movement and overall score is shown in Table 1. Other manual components such as palm orientation or location are not yet assessed (these parameters have since been implemented or are being implemented but were not yet fully developed at the time of data collection).

In addition to the scores, feedback is provided visually using the reference-video's skeleton overlaid with the learner-video's skeleton, whereby incorrect handshapes are marked by a red circle and hand movements by a red line in the feedback video. This enables learners to see at which point during the production and for which aspect an error occurred and for how long it lasted. It can be argued that this type of feedback is in line with the LOA principles outlined above as learners need to engage with the information provided and thereby develop their evaluative expertise.

The features of the Web SMILE Demo can be modified to different domains of use. First and foremost, the Demo is designed to be a learning tool in that it should help sign language learners perform the manual a spects of individual signs correctly. However, by deactivating access to the reference video, the Demo can also be used as an assessment tool, for example to test learners' vocabulary knowledge at the end of acourse.

RESEARCH QUESTIONS

The aims of the current paper were to 1) study the extent to which automated scores from the Web SMILE Demo correspond to human raters' scores, and 2) gauge learners' feedback regarding their experience with the Demo and the automated scores they receive. The following research questions were investigated:

(1) To what extent do human expert ratings agree with automated ratings for the five aspects rated by the Web SMILE Demo (handshape and hand movement for both dominant and non-dominant hand and overall score)?
(2) To what extent do learners perceive the automated scores and the automated feedback generated through the Web SMILE Demo to be useful?

METHODS

Three types of data were collected to answer the research questions: (1) automated scores from the Web SMILE Demo based on learner performances, (2) human raters' scores based on the same learner performances, and (3) learner feedback questionnaire responses. In addition, all participants filled in a background questionnaire. The participants and data collection procedures are described below.

PARTICIPANTS

Sign language learners

As a first step, we recruited 16 participants to work with the Demo and receive automated scores. The participants were all sign language interpreting learners (DSGS and German) in their first year of a bachelor's interpreting degree at a University in Switzerland. They received study credit points for their participation in this project. All learners were female, between 22 and 46 years old (*M* = 34.25, *SD* = 8.43), and either from Switzerland (*N* = 13), Liechtenstein (*N* = 2), or Germany (*N* = 1). In terms of their previous education, six learners had completed a professional certification (an apprenticeship or other professional training), eight learners had completed a degree in tertiary education (at a university or college), and two learners had completed both. German was the first language of all except one learner, whose first language was Romansh (a Romance language spoken in some eastern parts of Switzerland), however this learner indicated that she mostly used German in her daily life. None of the learners had hearing impairments, but three learners had a deaf partner or a deaf person in their extended family.

As part of the entry requirements for the interpreting degree, the learners needed to provide a DSGS course certificate, and we asked them which course they had completed. Ten learners had completed a CEFR (Common European Framework of Reference for Languages) A1-level course and 4 learners a CEFR A2-level course. We also asked the learners to self-rate their DSGS proficiency using the Language Experience and Proficiency Questionnaire (LEAP-Q, see Marian et al., 2007). Learners had to indicate their receptive and productive DSGS proficiency separately on an 11-point scale (where 0 = no proficiency and 10 = perfect proficiency). Learners rated themselves between 2 and 7 for reception (*M* = 3.25, *SD* = 1.44) and between 1 and 6 for production (*M* = 3.06, *SD* = 1.57).

Sign language raters

In addition to the sign language learners, we also recruited 14 expert raters for the study. The raters were all training to be DSGS teachers of deaf or hearing-impaired children and they were in the first year of a teacher training course at a University in Switzerland. The raters were compensated monetarily for their participation in the study. Ten raters were female and four were male. They were between 21 and 59 years old (*M* = 39.21, *SD* = 10.85), and either from Switzerland (*N* = 12), Brazil (*N* = 1), or Eritrea (*N* = 1); the two raters from Brazil and Eritrea had lived in Switzerland for 9 years. Prior to their teaching degree, 11 raters had completed a professional certification (an apprenticeship or other professional training), 1 rater had completed a degree in tertiary education (at a university or college), and 2 raters had completed both. All raters were deaf (*N* = 12) or severely hearing impaired (*N* = 2).

The raters had different DSGS backgrounds. Five raters acquired DSGS from birth as an L1, while the remaining nine raters picked up DSGS later in life and had learned it for 9 to 32 years (*M* = 21.13, *SD* = 7.51, with one missing response). One rater had learned Eritrean Sign Language as an L1 before acquiring DSGS, while 8 raters first acquired German (*N* = 7) or Albanian (*N*=1) as an L1. This is a common phenomenon among deaf people, who are often first exposed to a spoken language before acquiring a sign language (see Humphries et al., 2016). On the LEAP-Q, ten raters indicated that DSGS was the language they were most proficient in out of all the languages they knew, two raters put DSGS in second place, one rater in third, and one rater in fourth. This last rater, however, used two of the languages they were more proficient in only for writing. The raters self-rated their DSGS proficiency between 7 and 10 for reception (*M* = 8.93, *SD* = 1.00) and between 6 and 10 for production (*M* = 8.21, *SD* = 1.37), so they were all proficient to highly proficient users of DSGS.

DATA COLLECTION PROCEDURE

Learner performances

The first part of the data collection involved the learners performing the 16 lexemes through the Web SMILE Demo from home via their personal computers. In preparation for this, the learners first each booked a one-hour timeslot so we could monitor the Demo performance remotely for each learner and assist them in case of technical or other issues.

One week before their timeslot the learners received email instructions, which included software requirements (the Demo currently only runs on Google Chrome, so learners were asked to install the browser prior to data collection) and room setup (they were asked to be in a quiet room with a neutral background and stable internet connection). The learners also received access to the "How it works" section of the Demo and they were asked to familiarize themselves with it before their booked timeslot, to save time on the day of data collection. In addition, in the email we told the learners that they would have to produce each lexeme three times: the first time without looking at the reference video, the second time with looking at the reference video and to the best of their ability, and the third time with a deliberate mistake in their production such as incorrect handshape or hand movement. This enabled us to systematically investigate the extent to which the Web SMILE Demo detected errors as compared to the human raters. Feedback was provided to the learners after each recording.

The learners received another email one hour before their booked time slot, which included login information (username and password), a phone number and email address to contact us in case they needed assistance, and the link to the learner feedback questionnaire. During the data collection, the learners could choose in which order they wanted to produce the lexemes.

Learner feedback questionnaire

After the learners completed the Demo (i.e., after they produced each of the 16 lexemes three times) they filled out a feedback questionnaire via LimeSurvey. The questionnaire was piloted with three participants. After the piloting phase, we slightly changed the wording of three questions. The final version of the questionnaire included four main sections and a total of 22 questions and took about 10 minutes to complete. Sixteen questions were formulated as statements to which the learners had to indicate their level agreement on a Likert-scale (from "1 = fully agree" to "6 = fully disagree"), the remaining six questions were open-ended. In the first section, learners were asked seven Likert-scale and two open-ended questions about their experience with the Demo, such as clarity of instructions and technical difficulties. The second section included two Likert-scale questions and one open-ended question about their experienced difficulties in producing the lexemes. In the third section, we asked learners in six Likert-scale questions how relevant the skills practiced through the Demo were for their learning of DSGS, how interesting they found the Demo, and whether they enjoyed working on the tasks. Finally, the last section included one Likert-scale question and three open-ended questions about the quality of feedback the learners received through the Demo and whether they had any suggestions on how to improve the system. All learners answered the questions in the same order. The questionnaire was administered in German and is also available in English at http://tinyurl.com/2d8ksvah.

Rater training and rating form

After recruiting the 14 raters, we first trained them during two three-hour workshops as part of their teacher training at the university. For each performance, the raters scored the five different aspects according to the same scheme as used in the Demo: handshape individually for both hands (as either correct or incorrect), hand movement individually for both hands (on a scale from 1 to 4), and overall score (on a scale from 0 to 4, see Table 1 above). In the first workshop, we introduced the rating procedure and rating scale and asked them to familiarise themselves with it. Following this introduction, the raters rated an example performance on all five aspects during the workshop, and the results were compared and discussed. For the overall score, the raters were trained to base the ratings on their overall impression of the handshape and hand movement of both hands, as conveyed in the rating scale descriptors. After the initial training workshop, the raters received an Excel rating form including video links to three more example performances and they were asked to rate these on all aspects individually at home. During the next workshop one week later, the ratings were again compared and discussed as a group, to ensure that there was a mutual understanding of the scale descriptors and the procedure. In the course of the workshops, we also refined the rating instructions and changed the Excel rating form slightly to make them clearer.

After training, we sent the individual Excel rating forms including detailed instructions to all raters and asked them to complete the forms within one month. The final version of the rating forms included the gloss (meaning of the lexeme) and video link for each performance, as well as individual columns for the five aspects. The raters scored each aspect by marking the relevant cells with an *X*. The scale descriptors and rating instructions were also included in the rating form. An example rating form (translated from German to English) is included in Appendix 2.

Rating design

The 768 performances (16 learners × 16 lexemes × 3 attempts) were scored by the 14 raters on the five different aspects in an overlapping design, whereby all raters scored performances by all learners, on all lexemes, and for each of the three attempts. We ran a Python script to produce the 14 individual rating forms to ensure that the performances were allocated equally across raters in terms of the three different variables (learners, lexemes, and attempts). The script also controlled for the order of performances (e.g., if a particular learner, lexeme, or attempt would consistently have been rated last by many of the raters, this might have impacted the validity of the findings; controlling for the order was therefore important). In addition, the design included 96 anchor performances which were scored by all raters. The anchor performances were distributed equally across the three variables, and the position of the anchor videos within each rating form was also controlled for in the Python script to avoid potential ordering effects. Due to this research design, each rater scored between 143 and 145 performances (47 to 49 unique performances + 96 anchor performances).

DATA ANALYSIS

Rater and learner performance data were analysed via many-facet Rasch measurement (MFRM; Linacre, 1994) using the software package Facets (Linacre, 2023). We constructed a 5-facet model:

(1) *Rater*. Raters were grouped into human (Raters 1–14) and automated (Rater 15).
(2) *Learner*. Although learner signing ability was not a focus of the present study, it was necessary to control for the base signing ability of the learners, as such ability would impact their performance on the first, unsupported attempt.
(3) *Lexeme*. The 16 lexemes included in the study.
(4) *Attempt*. The first attempt was unsupported; the second was a reproduction of the sign as shown to the learner in the video; the third included a deliberate error.
(5) *Criterion*. Dominant handshape and non-dominant handshape were rated dichotomously; dominant hand movement and non-dominant hand movement were rated via a 4-level rating scale, and overall sign accuracy was rated via a 5-level rating scale whereby the top 4 levels were the same as for hand movement.

Rater groupings were used to compare the severity of the automated rater to the human raters using Wald *t* statistics (see Eckes, 2015). Facets bias/interaction analyses were employed to investigate differences regarding Criterion, Lexeme, and Attempt.

The questionnaire data for the closed questions were analysed descriptively by calculating the median answer for each statement. Open-ended responses were first grouped according to the six open-ended questions:

(1) What was not clear in the instructions? (5 responses)
(2) Which technical problems occurred? (4 responses)
(3) Did you experience any particular difficulties when completing the exercise? If yes, please briefly explain which difficulties. (6 responses)
(4) Do you have any suggestions on how the feedback could be improved to better support you in learning DSGS? (7 responses)
(5) Do you have any suggestions for other formats to help you learn DSGS even better? (2 responses)
(6) Is there anything else you would like to tell us about this exercise (e.g. suggestions on how the exercise could be improved)? (3 responses)

In a second step, the first author coded the responses for common themes, following the six phases of thematic analysis suggested by Brown and Clarke (2006). Through this process, 24 out of 27 responses could be unambiguously assigned one theme, while the remaining three responses were assigned two themes. The five identified themes were:

(1) Unclear instructions (5 responses)
(2) Technical problems (4 responses)
(3) Incorrect Demo feedback (12 responses)
(4) Suggested test formats (4 responses)
(5) Positive learner feedback (5 responses)

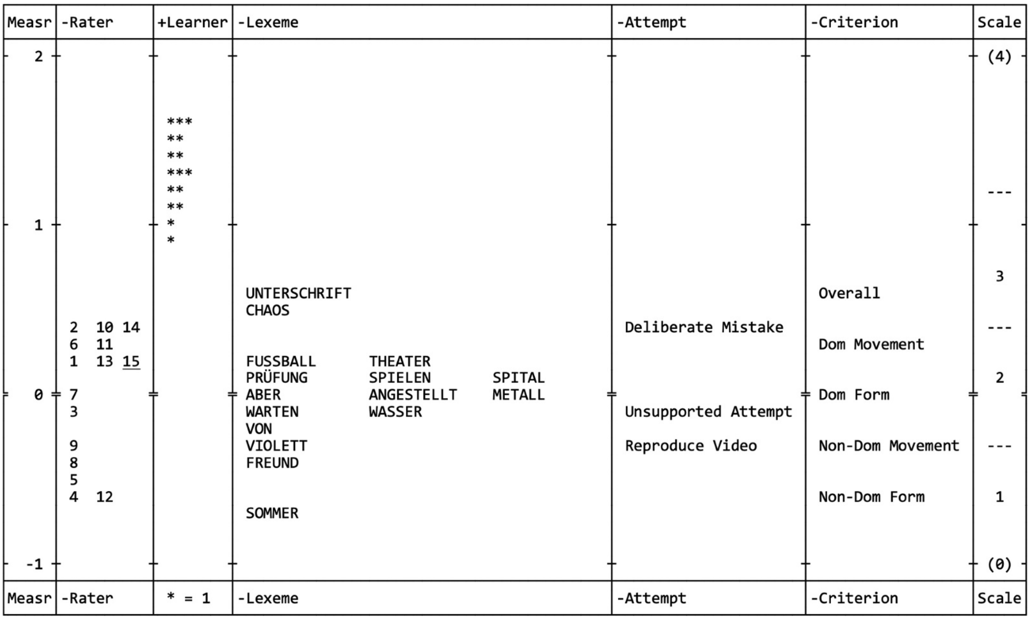
Table 2. MFRM Summary Statistics.

	Raters						
	All	Human	Automated	Learner	Lexeme	Attempt	Criterion
N	15	14	1	16	16	3	5
Measures							
Mean	0.00	−0.02	0.25	1.31	0.00	0.00	0.00
SD (sample)	0.38	0.39	0.00	0.21	0.31	0.37	0.49
SE	0.05	0.05	0.02	0.05	0.05	0.02	0.03
RMSE (sample)	0.05			0.05	0.05	0.02	0.04
Adjusted (True) *SD* (sample)	0.38			0.20	0.31	0.37	0.49
Infit *MS*							
Mean	1.04	1.03	1.10	1.04	1.05	1.04	1.03
SD (sample)	0.13	0.13	0.00	0.16	0.14	0.04	0.09
Outfit *MS*							
Mean	0.95	0.93	1.15	0.99	0.99	0.99	0.99
SD (sample)	0.12	0.10	0.00	0.17	0.24	0.10	0.10
Homogeneity index (χ2)	651.10			325.00	622.10	746.00	1021.20
df	14			15	15	2	4
p	0.00			0.00	0.00	0.00	0.00
Separation (sample)	7.15			4.40	6.45	18.56	13.36
Reliability of separation (sample)	0.98			0.95	0.98	1.00	0.99
Inter-rater reliability							
Observed exact agreement %	64.7						
Expected %	54.6						
Rasch κ	0.22						

RESULTS

MFRM measures and model fit

Before presenting results on the research questions, this section outlines MFRM measures to establish the extent to which our dataset cooperated with Rasch model requirements. Overall model fit was good, with Rasch measures accounting for 71.37% of the variance and the global chi square with a probability of .58. Rater fit indices were very close to the expected value of 1, with small standard deviations, indicating uniformity of fit. Rater Infit values ranged from .83 through 1.29, and Outfit values ranged from .72 through 1.16, indicating no serious departures from fit to the Rasch model (Wright & Linacre, 1994). A Rasch κ coefficient of 0.22 on the rater facet indicates a moderate level of rater agreement. Raters could be separated into seven distinct levels of severity. As expected, learners also had a range of abilities, being divisible into four distinct levels. The learners' overall ability was high compared to the difficulty of the lexemes, which was expected as most of the signs in the Demo were familiar to them, so their first attempt was rated highly. In addition, the second attempt was rated highly as learners were encouraged to produce the signs as best as they could modelling the reference video. Lexemes could be separated into six levels. The three attempts had a very wide range of difficulty, as would be hoped, given the inclusion of a "deliberate mistake" attempt. The Criterion facet also had a wide range of difficulty, ranging from the non-dominant handshape with a logit of – 0.64 through the overall criterion with a logit value of 0.58, representing 13 distinct levels of difficulty. All reliability of separation coefficients were 0.95 or higher. See Table 2 and Figure 1 for a summary of measures and model fit.



```
|Measr|-Rater   |+Learner|-Lexeme                                   |-Attempt            |-Criterion         |Scale|
|   2 +         +        +                                           +                    +                   + (4) +
|     |         |  ***   |                                           |                    |                   |     |
|     |         |  **    |                                           |                    |                   |     |
|     |         |  **    |                                           |                    |                   |     |
|     |         |  ***   |                                           |                    |                   |     |
|     |         |  **    |                                           |                    |                   | --- |
|     |         |  **    |                                           |                    |                   |     |
|   1 +         +  *     +                                           +                    +                   +     +
|     |         |  *     |                                           |                    |                   |     |
|     |         |        |                                           |                    |                   |  3  |
|     |         |        | UNTERSCHRIFT                              |                    | Overall           |     |
|     |         |        | CHAOS                                     |                    |                   |     |
|     | 2  10 14|        |                                           | Deliberate Mistake |                   | --- |
|     | 6  11   |        |                                           |                    | Dom Movement      |     |
|     | 1  13 15|        | FUSSBALL     THEATER                      |                    |                   |     |
|     |         |        | PRÜFUNG      SPIELEN      SPITAL          |                    |                   |  2  |
|   0 * 7       *        * ABER         ANGESTELLT   METALL          *                    * Dom Form          *     *
|     | 3       |        | WARTEN       WASSER                       | Unsupported Attempt|                   |     |
|     |         |        | VON                                       |                    |                   |     |
|     | 9       |        | VIOLETT                                   | Reproduce Video    | Non-Dom Movement  | --- |
|     | 8       |        | FREUND                                    |                    |                   |     |
|     | 5       |        |                                           |                    |                   |     |
|     | 4  12   |        |                                           |                    | Non-Dom Form      |  1  |
|     |         |        | SOMMER                                    |                    |                   |     |
|     |         |        |                                           |                    |                   |     |
|  -1 +         +        +                                           +                    +                   + (0) +
|Measr|-Rater   | * = 1  |-Lexeme                                    |-Attempt            |-Criterion         |Scale|
```

Figure 1. Wright Variable Map of MFRM Measures. Automated Rater is Underlined.

Comparisons of automated rater to human raters

Although the automated rater's severity was not the highest of the raters, the difference between its severity (0.25 logits) and the average of the human raters (–0.02 logits) was significant [$t(4553) = 5.01$; $p < .00001$], although the effect size was small ($d = 0.15$). A summary of significant contrasts between the human raters' biased severity measures and those of the automated rater for Criterion, Lexeme, and Attempt can be seen in Table 3. Negative contrast values indicate that the automated rater was more severe.

Bias/interaction analyses found the largest average contrast among the significant differences between the human raters' and the automated rater's biased severities to be on the Lexeme facet, with 24 lexeme comparisons exhibiting contrasts with large Cohen's *d*effect sizes according to the Plonsky and Oswald (2014) two-sample thresholds for second language a cquisition research (.60 = small, 1.00 = medium, 1.40 = large). The lexemes with the most significant contrasts were SPITAL (*hospital* in English) and WASSER (*water*), although the largest contrasts were observed with WASSER (1.38 logits) and VIOLETT (*purple*; 1.32 logits).

The second largest average contrast was found on the Criterion facet, with the largest (1.13 logits) and most (12) contrasts appearing on the Non-Dominant Form criterion, although none of the effect sizes were large.

Finally, the facet with the smallest average contrast was that of Attempt, of which the attempt with the largest contrast and highest number of differences with meaningful effect sizes was that of Reproduce Video, wherein learners signed immediately after viewing a video of the citation form of each sign. However, there was roughly an equal number of significant differences between humans and the automated rater for the Reproduce Video (8) and Deliberate Mistake (9) attempts.

Table 3. Summary of Significant Contrasts between Human Raters and Automated Rater.

			Avg. Contrast		Cohen's *d**		
	N	%	Raw	Abs.	S	M	L
Criterion (TOTAL)	36	100%	−0.43	0.69	20	5	0
Dom Form	7	19%	−0.75	0.75	4	1	0
Non-Dom Form	12	33%	−1.13	1.13	8	2	0
Dom Movement	2	6%	−0.24	0.24	0	0	0
Non-Dom Movement	9	25%	0.44	0.44	5	1	0
Overall	6	17%	−0.02	0.30	3	1	0
Lexeme (TOTAL)	78	100%	−0.04	0.82	26	30	24
ABER	3	4%	0.53	0.53	1	2	0
ANGESTELLT	4	5%	−0.31	0.57	1	3	0
CHAOS	8	10%	−0.54	0.54	3	5	0
FREUND	5	6%	0.60	0.60	3	2	0
FUSSBALL	1	1%	−0.75	0.75	2	0	0
METALL	5	6%	0.62	0.62	1	4	0
PRÜFUNG	1	1%	−0.56	0.56	2	0	0
SOMMER	1	1%	0.56	0.56	1	0	0
SPIELEN	2	3%	0.03	0.47	2	0	0
SPITAL	11	14%	0.80	0.80	2	4	5
THEATER	7	9%	0.81	0.81	1	2	4
UNTERSCHRIFT	9	12%	0.83	0.83	0	2	7
VIOLETT	3	4%	−1.32	1.32	2	1	0
WARTEN	4	5%	−0.56	0.56	4	0	0
WASSER	14	18%	−1.38	1.38	1	5	8
Attempt (TOTAL)	20	100%	−0.12	0.39	7	1	0
Unsupported Attempt	3	15%	−0.11	0.27	1	0	0
Reproduce Video	8	40%	−0.57	0.57	5	0	0
Deliberate Error	9	45%	0.27	0.27	1	1	0

*Size thresholds according to Plonsky and Oswald (2014) two-sample recommendations.

Overall, although the automated rater was significantly harsher than the average of the human raters, and several significant contrasts between it and individual human raters were observed, the automated rater was not significantly different from the human raters nearest to it in severity (1, 6, 11, 13). This result, coupled with its InfitMS value of 1.10, indicates that the automated rater scores similarly to a human rater of slightly-above-average severity.

Learner feedback

Learner responses to the Likert-scale questionnaire statements are shown in Figure 2. As shown at the top the figure, learners agreed or fully agreed with the statements about administration included in the questionnaire, indicating that they were generally satisfied with that aspect. There were minimal misunderstandings with regards to the instructions, however these did not impact the learners' performance: The open responses related to unclear instructions (Theme 1) showed, for example, that one learner was unsure whether they had finished the exercise because the Demo did not include a message after the learners had completed all signs three times. Technical problems (Theme 2) were minor and mostly related to long processing times, which was likely caused by slow internet connection. In addition, learners did not think that the exercise was very difficult, and they did not find it difficult to use the Demo.

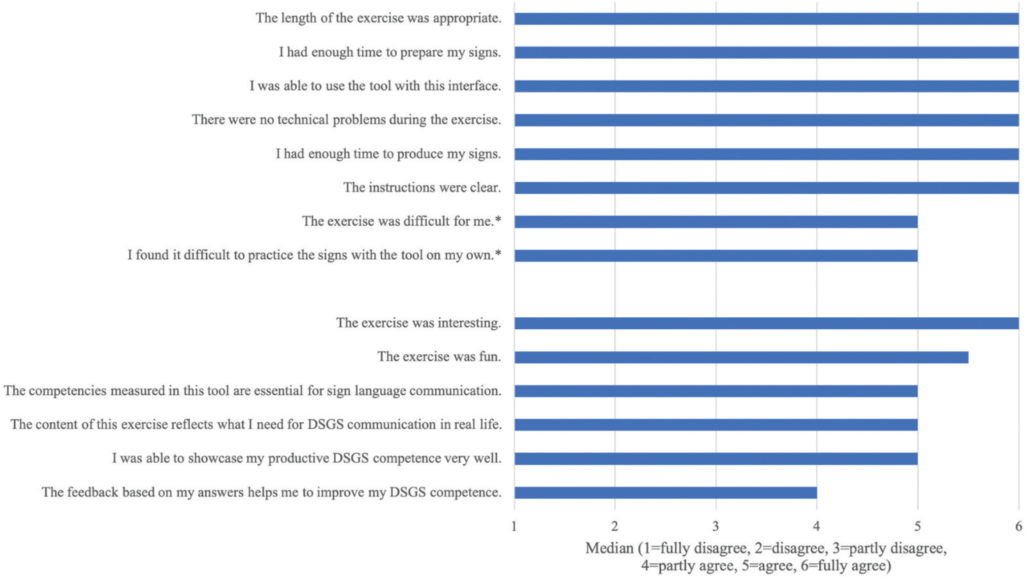


Figure 2. Student Questionnaire Responses. *The two statements on difficulty were reverse-coded.

As shown at the bottom of Figure 2, the learners thought that the exercise was interesting and fun and that the competencies and contents measured were important for their sign language learning. This was confirmed by the analysis of the open responses, five of which were coded as “Positive learner feedback” (Theme 5) and included statements such as “Great software, thank you!”. However, several learners felt that they could not showcase their productive DSGS-competence very well, although the group as a whole agreed to this statement ($Mdn = 5$). One reason for this may be that many learners may have already known many of the lexemes and may therefore not have been able to show their full productive DSGS-proficiency. Some of the learners also suggested other test format for measuring productive signing in their open responses (Theme 4), for example “sentence-level assessments”.

The statement “The feedback helps me to improve my DSGS-competence” received the lowest level of agreement, with six learners partly disagreeing and one learner fully disagreeing ($Mdn = 4$ “partly agree”). The reasons why this statement received a relatively low level of agreement become clear when looking at the learners’ open answers: Incorrect Demo feedback (Theme 3) was the most common theme across all responses, with 6 out of 16 students mentioning in 12 responses that they sometimes did not understand why the Demo rated certain aspects as incorrect. Exemplary comments were:

> Learner 1: *Sometimes I couldn’t make out why the program declared something as correct or incorrect. For example, with [“von”] I thought it must be because of small discrepancies in the recording angle.*
>
> Learner 5: *[. . .] I was irritated that I sometimes received the message “The target sign was incorrect” when I definitely did not see an error.*

> Learner 12: *Sometimes it was not clear for me at all which mistakes I had made or where the differences between my video and the reference video were, so I could not make any improvements.*

> Learner 16: *In isolated cases it was not clear for me what the mistake was. For example, when I put down my hand [at the end of a sign], this was also rated.*

DISCUSSION AND CONCLUSION

This study investigated the extent to which human expert raters agree with automated ratings of the Web SMILE Demo, an application designed for learning and automatically assessing the production of individual lexemes in sign languages. We also studied learners' perception of the automated scores and the automated feedback generated through the Demo by means of a questionnaire. The study was conducted in the context of DSGS.

The results indicate that the Web SMILE Demo was more severe than an average human rater, however it showed a high level of agreement with the human raters nearest to it in severity. In terms of the different rating criteria, the automated ratings on the non-dominant hand showed larger discrepancies with the human ratings than the ratings on the dominant hand (58% of comparisons with significant differences were observed for the non-dominant hand, compared to 25% for the dominant hand). Subsequent analyses indicated that the automated rater picked up on small movements of the non-dominant hand even for one-handed signs and rated them lower, whereas human raters would likely have considered small movements as acceptable. This finding is in line with comparisons of human and automated ratings for assessing writing or speaking performances in spoken languages, where high levels of overall agreement between humans and machines are usually reported, despite some deviations in severity for certain aspects (for assessing writing see e.g. Chan et al., 2023; Correnti et al., 2020; Hussein et al., 2019; for assessing speaking see e.g.; Xi et al., 2012; Xu et al., 2021). Thus, the automated ratings reported through the Web SMILE Demo can be said to reflect human expert ratings well, although they were slightly more severe than an average human rater, particularly for rating handshape and hand movement of the non-dominant hand. Apart from displaying slightly-above average rater severity, the many-facet Rasch model also showed that the Web SMILE Demo rated consistently, with an InfitMS value of 1.10. This indicates a good level of intra-rater reliability.

Despite good agreement between the automated system with (slightly severe) human expert raters, as well as high intra-rater reliability, learner feedback questionnaire data indicated that some of the automated ratings were confusing for the learners. It was not always clear to learners why they received a low score on certain aspects of their performance, when they could not see any obvious errors in their production. Thus, while we found good human-machine and machine-internal agreement at an overall statistical level, the automated system occasionally generated ratings which impacted the quality of feedback at the level of individual lexemes for individual learners. These ratings clearly confused some learners and frustrated them, as they sometimes received a low score from the Demo despite no obvious discrepancies between their performance and the one in the reference video. Some of these ratings may be due to the severe automated ratings of the non-dominant hand criteria for one-handed signs outlined above. Like the human expert raters, the learners would not have considered small movements of the non-dominant hand as problematic, and therefore were confused by the severe ratings. Thus, from a LOA perspective, the scores generated through the Demo need to be aligned to ensure that the automated ratings consistently correspond to the performances, so that learners receive useful feedback to improve their learning.

We would also argue that deviations of this kind have the potential to sow distrust in the automated feedback, despite a high degree of accuracy of automated ratings on an overall statistical level. Although such ratings were not overly common, they did occur for over a third of learners in our experiments and thereby impacted these learners' impression on the usefulness of the tool as a learning and assessment platform. To our knowledge, previous research on automated scoring in sign languages has not addressed this issue. This may have to do with the fact that automated sign language scoring is only just emerging and has so far only been used in a limited number of contexts (Chai et al., 2017; Ellis et al., 2016; Liu et al., 2024; Phan et al., 2018). Based on our data, we would argue that automated feedback alone can sometimes lead to misunderstandings in situations when the system produces ratings for specific aspects of a performance that do not reflect the actual performance. A disadvantage of automated systems is that they usually do not outline to the user how they reached a certain score. Although Machine Learning models are transparent and can be explained, the underlying algorithms are often not comprehendible for lay-people. One option to address this shortcoming may be to give learners the opportunity to ask for (human) clarification in case the automated feedback is obviously incorrect. Once the system has been comprehensively validated to ensure that incorrect ratings do not occur, human clarifications may no longer be necessary.

Apart from some of the automated ratings impacting the overall quality of feedback, learners generally enjoyed working with the Web SMILE Demo. They thought that the automated system was interesting and fun, and that it measured competencies which the learners needed for sign language communication. This indicates that the tasks were authentic and elicit competencies that mirror those needed in real-life settings, which is another necessary prerequisite for successful LOA (Carless et al., 2006). In addition, for many of the signs and in many instances, the Demo provided useful feedback on the correctness of handshape and hand movement for both dominant and non-dominant hand. The Demo was successful in helping learners practice individual signs in DSGS, and to perform the manual aspects of these signs accurately. The fact that the learners provided insightful comments on the feedback they received also shows that they developed and applied their evaluative expertise, which addresses the third key principle of LOA (Carless, 2015). Their comments showed that improvements need to be made on rating accuracy, for example regarding the automated ratings of the non-dominant hand, to reduce the number of overly severe ratings.

FUTURE WORK

Based on the data collected as part of this study, the Web SMILE Demo will be further developed and refined. The rating data gathered in the current study will be used to train the automated system further and to align it more closely with human raters of average rating severity. To make the feedback even more user-friendly, more precise video feedback including skeletal renderings for each hand will be implemented and detailed information on palm orientation and location will be added. In addition, the model for the recognition of the individual aspects will be refined by including additional lexemes, both one-handed and two-handed, in the next version of the system.

ETHICS STATEMENTS AND DECLARATIONS

Prior to the study, ethical approval was obtained from the first author's university and all participants gave informed written consent to take part in the research.

The research data can be obtained by contacting the first author.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung [193686].

ORCID

Franz Holzknecht http://orcid.org/0000-0002-1218-2062
Sandrine Tornay http://orcid.org/0000-0002-9273-9325
Alessia Battisti http://orcid.org/0000-0002-1696-6921
Aaron Olaf Batty http://orcid.org/0000-0002-2760-4918
Katja Tissi http://orcid.org/0000-0002-5059-8210
Tobias Haug http://orcid.org/0000-0002-8713-1163
Sarah Ebling http://orcid.org/0000-0001-6511-5085

References

Bayley, R., Schembri, A., & Lucas, C. (2015). Variation and change in sign languages. In A. Schembri & C. Lucas (Eds.), *Sociolinguistics and deaf communities* (pp. 61–94). Cambridge University Press. https://doi.org/10.1017/CBO9781107280298.004

Brown, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carless, D. (2015). Exploring learning-oriented assessment processes. *Higher Education*, *69*(6), 963–976. https://doi.org/10.1007/s10734-014-9816-z

Carless, D., Joughin, G., & Liu, N.-F. (2006). *How assessment supports learning: Learning-oriented assessment in action*. Hong Kong University Press. https://doi.org/10.5790/hongkong/9789622098237.001.0001

Chai, X., Liu, Z., Li, Y., Yin, F., & Chen, X. (2017). SignInstructor: An effective tool for sign language vocabulary learning. *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 900–905). Nanjing, China. https://doi.org/10.1109/ACPR.2017.130

Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. *Language Testing*, *40*(1), 61–85. https://doi.org/10.1177/02655322221076025

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., & Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, *55*(3), 493–520. https://doi.org/10.1002/rrq.281

Davis, L., & Papageorgiou, S. (2021). Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral a cademic English. *Assessment in Education Principles, Policy & Practice*, *28*(4), 437–455. https://doi.org/10.1080/0969594X.2021.1979466

Ebling, S., Camgoz, N. C., Braem ,P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., & Magimai-Doss, M. (2018). SMILE Swiss German sign language dataset. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.). *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)* (pp. 4221–4229). Miyazaki, Japan: European Language Resources Association (ELRA).

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

Ellis, K., Fisher, J., Willoughby, L., & Barca, J. C. (2016). *A design science exploration of a visual-spatial learning system with feedback*. Paper presented at the Australasian Conference on Information Systems 2015, Adelaide. https://doi.org/10.48550/ARXIV.1606.03509

Fenlon, J., & Hochgesang, J. A. (Eds.). (2022). *Signed language corpora*. Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc

Hanke, T., & Fenlon, J. (2022). Creating corpora: Data collection. In J. Fenlon & J. A. Hochgesang (Eds.), *Signed language corpora* (pp. 18–45). Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.7

Haug, T., & Holzknecht, F. (2020). Using automatic sign language recognition for sentence-level assessment of Swiss German Sign Language – SMILE-II. *Paper presented at the 2020 Language Testing Research Colloquium*, online.

Haug, T., Mann, W., & Holzknecht, F. (2023). The use of technology in sign language testing: Results of a pre-pandemic survey. *Sign Language Studies*, *23*(2), 243–281. https://doi.org/10.1353/sls.2023.0003

Humphries, T., Kushalnagar, P., Mathur, G., Napoli, D. J., Padden, C., Rathmann, C., & Smith, S. (2016). Language choices for deaf infants: Advice for parents regarding sign languages. *Clinical Pediatrics*, *55*(6), 513–517. https://doi.org/10.1177/0009922815616891

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: Aliterature review. *Peer J Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Jin, Y. (2023). Test-taker insights for language assessment policies and practices. *Language Testing*, *40*(1), 193–203. https://doi.org/10.1177/02655322221117136

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

Linacre, J. M. (2023). *Facets (3.85.1)*. Winsteps.com. http://www.winsteps.com/facets.htm

Liu, Z., Pang, L., & Qi, X. (2024). MEN: Mutual enhancement networks for sign language recognition and education. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 311–325. https://doi.org/10.1109/TNNLS.2022.3174031

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Phan, H. D., Ellis, K., & Dorin, A. (2018). MIC, an interactive sign language teaching system. In D. McKay, J. Waycott, A. Morrison, J. -H. -J. Choi, A. Lugmayr, M. Billinghurst, R. Kelly, G. Buchanan, & D. Stevenson (Eds.). *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (pp. 544–547). Melbourne, Australia. https://doi.org/10.1145/3292147.3292237.

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

Schembri, A., & Cormier, K. (2022). Signed language corpora: Future directions. In J. Fenlon & J. A. Hochgesang (Eds.), *Signed language corpora* (pp. 196–220). Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.12

Tornay, S., Camgoz, N. C., Bowden, R., & Magimai Doss, M. (2020). A phonology-based approach for isolated sign production assessment in sign language. *Companion Publication of the 2020 International Conference on Multimodal Interaction* (pp. 102–106). Utrecht, The Netherlands. https://doi.org/10.1145/3395035.3425251

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). De Gruyter. https://doi.org/10.1515/9781614513827-023

Watkins, F., & Thompson, R. L. (2017). The relationship between sign production and sign comprehension: What handedness reveals. *Cognition*, *164*, 144–149. https://doi.org/10.1016/j.cognition.2017.03.019

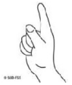
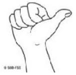
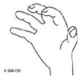
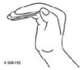
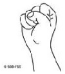
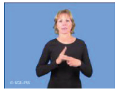
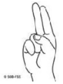
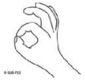
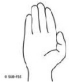
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing*, *29*(3), 371–394. https://doi.org/10.1177/0265532211425673

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education Principles, Policy & Practice*, *28*(4), 411–436. https://doi.org/10.1080/0969594X.2021.1979467

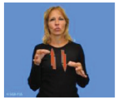
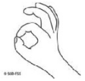
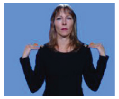
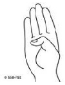
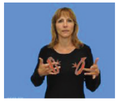
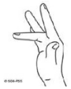
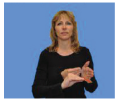
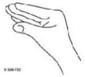
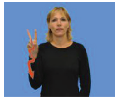
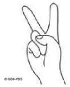
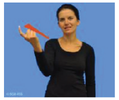
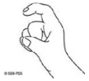
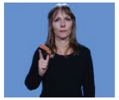
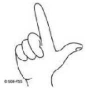
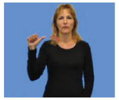
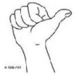
Appendices

Appendix 1 Details about the 16 lexemes

Gloss	Image	Handedness	Handshape dominant hand	Movement	Links to the images and videos from the SGB-FSS* online lexicon:
ABER but		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/116541/aber
ANGESTELLTE employee		Two hands			https://www.sgb-fss.ch/signsuisse/lexikon/123654/angestellte
CHAOS chaos		Two hands		asymmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/128590/chaos
FREUND friend		Two hands			https://www.sgb-fss.ch/signsuisse/lexikon/112906/freunde
FUSSBALL soccer		Two hands		asymmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/114842/fussball
METALL metal		Two hands		symmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/130034/metall
PRÜFUNG exam		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/124295/pruefung
SOMMER summer		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/114804/sommer

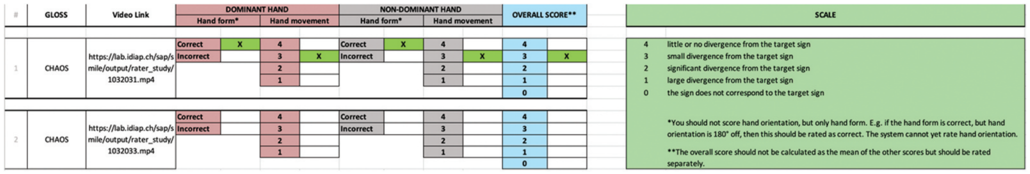
(*Continued*)

(Continued).

Gloss	Image	Handedness	Handshape dominant hand	Movement	Links to the images and videos from the SGB-FSS* online lexicon:
SPIELEN to play		Two hands		asymmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/124539/spielen
SPITAL hospital		Two hands		symmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/126431/spital
THEATER theater		Two hands		asymmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/123605/theater
UNTERSCHRIFT signature		Two hands		asymmetrical	https://www.sgb-fss.ch/signsuisse/lexikon/113473/unterschrift
VIOLETT violet		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/117453/violett
VON from		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/112632/von
WARTEN to wait		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/126727/warten
WASSER water		One hand			https://www.sgb-fss.ch/signsuisse/lexikon/125816/wasser

*Swiss Association of the Deaf (Schweizerischer Gehörlosenbund, Fédération Suisse des Sourds, Federazione Svizzera dei Sordi).

Appendix 2 Exemplary rating form



#	GLOSS	Video Link	DOMINANT HAND				NON-DOMINANT HAND				OVERALL SCORE**	
			Hand form*		Hand movement		Hand form*		Hand movement			
1	CHAOS	https://lab.idiap.ch/sap/smile/output/rater_study/1032031.mp4	Correct	X	4		Correct	X	4		4	
			Incorrect		3	X	Incorrect		3	X	3	X
					2				2		2	
					1				1		1	
											0	
2	CHAOS	https://lab.idiap.ch/sap/smile/output/rater_study/1032033.mp4	Correct		4		Correct		4		4	
			Incorrect		3		Incorrect		3		3	
					2				2		2	
					1				1		1	
											0	

SCALE	
4	little or no divergence from the target sign
3	small divergence from the target sign
2	significant divergence from the target sign
1	large divergence from the target sign
0	the sign does not correspond to the target sign

*You should not score hand orientation, but only hand form. E.g. if the hand form is correct, but hand orientation is 180° off, then this should be rated as correct. The system cannot yet rate hand orientation.

**The overall score should not be calculated as the mean of the other scores but should be rated separately.

Note: The first two performances are shown. Only the first performance has been scored (as indicated by the green cells marked with "X").