Evaluation of a Sign Language Avatar on Comprehensibility, User Experience & Acceptability

Fenya Wasserroth
Technische Universität Berlin
Berlin, Germany
wasserroth@campus.tu-berlin.de

Eleftherios Avramidis
German Research Center for AI
Berlin, Germany
eleftherios.avramidis@dfki.de

Vera Czehmann
German Research Center for AI
Berlin, Germany
Technische Universität Berlin
Berlin, Germany
vera.czehmann@dfki.de

Tanja Kojic
Technische Universität Berlin
Berlin, Germany
tanja.kojic@tu-berlin.de

Fabrizio Nunnari
German Research Center for AI
Saarbrücken, Germany
fabrizio.nunnari@dfki.de

Sebastian Möller
Technische Universität Berlin
Berlin, Germany
German Research Center for AI
Berlin, Germany
sebastian.moeller@tu-berlin.de

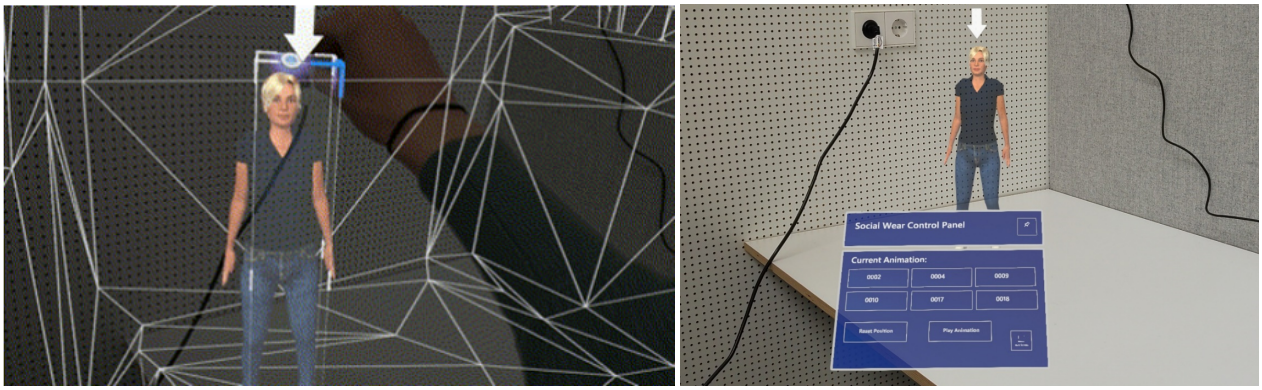


Figure 1: Avatar modification example

Abstract

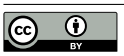
This paper presents an investigation into the impact of adding adjustment features to an existing sign language (SL) avatar on a Microsoft Hololens 2 device. Through a detailed analysis of interactions of expert German Sign Language (DGS) users with both adjustable and non-adjustable avatars in a specific use case, this study identifies the key factors influencing the comprehensibility, the user experience (UX), and the acceptability of such a system. Despite user preference for adjustable settings, no significant improvements in UX or comprehensibility were observed, which remained at low levels, amid missing SL elements (mouthings and facial expressions) and implementation issues (indistinct hand shapes, lack of feedback and menu positioning). Hedonic quality was rated higher than pragmatic quality, indicating that users found the system more emotionally or aesthetically pleasing than functionally useful. Stress levels were higher for the adjustable avatar, reflecting lower performance, greater effort and more frustration. Additionally, concerns were raised about whether the Hololens adjustment gestures are intuitive and easy to familiarise oneself with. While acceptability of the concept of adjustability was generally positive, it was strongly dependent on usability and animation quality. This study highlights that personalisation alone is insufficient, and that SL avatars must be comprehensible by default. Key recommendations include enhancing mouthing and facial animation, improving interaction interfaces, and applying participatory design.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Applied computing** → *Language translation*; • **Human-centered computing** → **Accessibility systems and tools**; **Mixed / augmented reality**; *Natural language interfaces*; **User studies**; **Usability testing**.

Keywords

sign language, avatar, UX, acceptability, comprehensibility



IVA Adjunct '25, Berlin, Germany

ACM ISBN 979-8-4007-1996-7/2025/09
https://doi.org/10.1145/3742886.3756719

ACM Reference Format:
Fenya Wasserroth, Eleftherios Avramidis, Vera Czehmann, Tanja Kojic, Fabrizio Nunnari, and Sebastian Möller. 2025. Evaluation of a Sign Language Avatar on Comprehensibility, User Experience & Acceptability. In *ACM International Conference on Intelligent Virtual Agents (IVA Adjunct '25), September 16–19, 2025, Berlin, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3742886.3756719

1 Introduction

Over 70 million people worldwide are deaf and use more than 200 national sign languages (SLs) [43]. For many, SL is their first and preferred language, and written text remains a less accessible second language. SL interpreters thus support communication between hearing individuals and deaf people. According to the Federal Association of Sign Language Interpreters in Germany, around 850 interpreters serve approximately 80,000 hearing-impaired individuals in Germany [5]. As a result, there are everyday situations where no interpreter is available, complicating communication and limiting access—for example, to train announcements.

SL avatars could help improve accessibility, making research in this area crucial. While one might argue displaying text subtitles on AR glasses could be sufficient, this overlooks the cultural and linguistic significance of SLs. Prioritising text over SL in assistive technologies risks further marginalising SL users and perpetuating a hearing-centric view of communication. If successful, communication barriers between deaf and hearing communities could be reduced long-term. These avatars are not intended to replace human interpreters, but to support the autonomy of deaf individuals in situations where interpreters are unavailable [18].

At the same time, *augmented reality* (AR) technology is evolving, finding applications in fields like surgery [15] and automotive displays. AR can embed artificial information into the real environment—an advantage that can benefit SL avatars [44]. In research led by hearing individuals, expertise from the deaf and hard-of-hearing community is particularly important [11].

This study examines how avatar adjustments—such as scalable size and adjustable spatial positioning—affect comprehensibility, user experience, and acceptability in an AR SL assistant. A user study with a mixed-method approach was conducted that combined quantitative data from standardised questionnaires with qualitative responses from an interview with the participant. The following research questions guide the investigation:

- **Comprehensibility**: Can adjustment options of a SL avatar improve the comprehensibility of the signed sentences?
- **User experience**: Do adjustment options contribute to a better user experience of a SL avatar?
- **Acceptability**: To what extent do the adjustable setting options contribute to the acceptability of a SL avatar? Are there differences in acceptability and, if so, can factors be identified that explain these differences?

2 Related work

2.1 Sign language avatars

Research in this area focuses on three core themes: acceptance, qualitative user requirements, and technical development.

Qualitative requirements include both the technical necessities of transmitting SL and the needs of the deaf and hard-of-hearing community. SL involves hand shapes, movement, location, facial expressions, and lip movements. These components must be animated and synchronised, and signs combined into fluid sentences [4, 18]. Facial expressions and lip movements are crucial for understanding and emotional expression [16, 19]. Presentation and control aspects are also being studied. Avatars may have comic-like or human-like appearances [3], though comic styles can be situationally appropriate [16]. More human-like avatars risk entering the "uncanny valley", reducing acceptance [6, 24]. Personalisation is valued, but customisations that improve comprehension—like size or speed—are considered more important than appearance [16, 30]. Technically, avatars must convert spoken language into signed output. This involves automated translation and/or avatar synthesis—challenging, because SL is neither linear nor written. Methods involving motion capture (MoCap) or glosses are common but limited [41]. Precise synchronisation of visual elements is required [3, 41].

The use of Augmented Reality for SL avatars has been suggested in order to allow deaf or hard of hearing people to see an interpreter next to the speaking person or other visual information simultaneously, without having to switch their gaze [22, 27, 28]. Many efforts have focused on educational use cases [1, 7, 17, 21, 36, 39, 45], most of them in prototypical stage with either no or only limited user evaluation. Among these, [45] contains a user study similar to ours, showing that users prefer a full-body vs. a half-body signing AR avatar, with the possibility to adjust its position.

The DFKI avatar for DGS [31], on which this work is based (Figure 1), is an animated AR SL interpreter currently running on Microsoft HoloLens 2. The version used for this experiment can sign six pre-recorded sentences related to public transport, developed within the AVASAG research project [32]. Users can start animations via a menu after a brief loading time. The avatar's size, 3D position, and orientation are adjustable. A visible frame appears when the user's hand is nearby. Users can scale the avatar by dragging a corner, and rotate it using side handles.

2.2 Comprehensibility

To investigate the understanding of natural language, the different concepts must first be defined and differentiated from one another. The following levels are distinguished: Understandability, comprehensibility, communicability and comprehension [25]. Understandability describes the lowest level, as just simply the ability to transmit the signal [25], and can apply to units of various sizes. Comprehensibility describes the identification of the signal based on its form [25]. It is influenced by the comprehensibility of individual units and by lexical, syntactic and semantic context [25]. When applied to SL, comprehensibility describes how well the content of a signed statement can be identified [8]. Communicability describes whether an utterance is understood in the way it was intended [25]. In addition to comprehensibility, this is influenced by meta-factors such as delay times of the signal. A successful communication process ends in understanding [25].

Various metrics are used to evaluate the quality of the animated SL avatars. Metrics from machine translation (MT) such as the word

error rate (WER), the BiLingual Evaluation Understudy (BLEU) or NIST, a further development of BLEU [14], are frequently used. These metrics, however, were designed to assess precision by comparing output to a human translation [14, 34]. While well established for machine text-to-text translations, there is no standardised method of evaluating text-to-SL translations [26]. Various forms of sign error rate (SER), such as the multiple reference sign error rate (mSER), are also used in the evaluation of SL [23]. All of these metrics measure comprehensibility, but do not yet provide any information about how comprehensible it is for users.

Martino et al. [10] assessed the comprehensibility of individual signs using an "isolated sign comprehensibility test" (ISIT), in which the animation is evaluated against the video of a SL interpreter. Nevertheless, Crowe et al. [8] report a lack of studies on the comprehensibility of SL from the recipient's perspective, which is an indicator that there is no standardised method. However, the importance of comprehensibility for SL avatars has been researched: especially in important areas of application, where errors in comprehensibility have critical consequences, full comprehensibility must be guaranteed regardless of the SL level [18]. Crowe et al. [8] investigated which factors influence the comprehensibility of SL. The use of grammatical features and the clarity of signs are considered to promote comprehensibility, whereas the exclusive use of the finger alphabet leads to lower comprehensibility.

On the question of what influence comprehensibility has on the acceptance and evaluation of SL avatars, Wolfe et al. [42] predict a very large influence, suggesting that judgement is dependent on the translation quality and the associated effort of comprehension. In a study with focus groups, however, it was observed that the participants did not differentiate between the pure translation quality and the animation [18]. Nevertheless, these results show that the comprehensibility of animated signs plays a key role in the research and development of SL avatars. In addition, the study by Smith and Nolan [38] suggests that the appearance of the avatar affects comprehensibility too. For example, a human-like avatar achieved a better comprehensibility rate than a comic-like one.

2.3 Acceptability

Acceptability describes the proportion of the target group that would actually use a SL avatar [25]. Therefore, only the preliminary stage of acceptability can be observed by the specified use in the context of the evaluation [2]. There is a wide range of models and approaches in the literature to describe factors of acceptability. A well-known model is the Technology Acceptance Model (TAM), which is based on perceived *usefulness* and *ease of use* [9]. [33] extended this to TAMSA, adding the factor *trust*, due to lower trust in avatars compared to human interpreters. TAMSA, however, was developed with Deaf individuals in Qatar exclusively, limiting generalisability [33]. Acceptability often depends on use context [16, 18]. Nielsen's model distinguishes social from practical acceptability [29]. Social acceptability is influenced by context, norms, and experience, while practical acceptability covers aspects like usefulness and reliability. Social acceptability improves when avatars are intended to complement interpreters in otherwise inaccessible contexts [18]. Independence and flexibility enhance perceived usefulness. Voluntary use is also key [16]. Individual background influences acceptance – people who learned SL early tend to show less interest in avatars [37].

Technical challenges must be addressed to meet user expectations. Including deaf individuals in the development process provides essential expertise and fosters trust [18, 35]. Engaging users in development-related focus groups has also been shown to enhance social acceptability. [16].

3 Methodology

3.1 Evaluation Design

The evaluation combined summative and formative aspects – while the implemented functions of the SL avatar were summatively assessed, the findings were also used formatively to support future development. Accordingly, practical testing of the SL avatar application on the Microsoft HoloLens 2 is a core component of the investigation. To address the research questions, a user study was designed with a focus on deaf and hard-of-hearing participants.

To analyze the differences identified in the research questions, the study compared two variants of the SL avatar—one fixed and one adjustable. Technically, both variants are identical. However, in the fixed version, the avatar was displayed in a default position and users were restricted from interacting with it via gesture-based adjustments.

By using a within-subject design, the study captured relative rather than absolute evaluations of the avatar —an important distinction given the current state of avatar development. In the adjustable version (A), participants could modify the avatar's size, orientation, and spatial positioning using gestures. In the fixed version (F), the avatar remained anchored within the user's field of vision, scaled to a legible size. The version presented first (either F or A) acts as an anchor for the comparative evaluation of the second condition.

To assess comprehensibility rather than overall comprehension, ideally, signs would be presented in isolation and without contextual cues, similar to the ISIT approach described by [10]. However, many signs—such as pointing or directional signs—derive their meaning from contextual cues within full sentences. Given that the avatar currently only supports playback of pre-stored sentences, the study instead compared comprehensibility across two conditions: one in which participants were given a described situational context (I), and one in which sentences were presented without any contextual information (II). This allowed for an exploration of whether contextual knowledge and prior expectations influence the comprehensibility of avatar-signed content.

This results in the test conditions reported in Table 1.

Test Condition	**Fixed Avatar (F)**	**Adjustable Avatar (A)**
Situational comprehensibility (I)	F-I	A-I
Pure Comprehensibility (II)	F-II	A-II

Table 1: Test conditions

In accordance with the previous description of the evaluation design, each participant tested both versions (F and A) of the SL avatar (see Figure 1 for a screenshot and Figure 2 for the procedure).

Half of the participants started by testing the fixed avatar (F), while the other half started with the adjustable avatar (A). The order in which participants encountered the situational comprehensibility condition (I) versus the pure comprehensibility condition (II) was also counterbalanced. This variation in sequence was intended to minimize learning and order effects. During the testing phases, objective methods were applied: comprehensibility was assessed using performance-based measures. Additionally, observational data on how participants interacted with the application were collected to provide insights into usability and user experience. Participants' visual perspectives were also recorded via screen capture, enabling retrospective, objective evaluation.

Following each test phase, subjective impressions of user experience and stress levels were collected using standardised questionnaires. These closed-format items facilitate comparison across participants. However, for the formative aspect of the evaluation, qualitative feedback was essential. Therefore, at the end of each session, participants were interviewed using open-ended questions. The combination of qualitative and quantitative data allowed for a comprehensive understanding of user perception, with a particular emphasis on how the system was understood [40].

Participants could choose whether they preferred the instructions and questionnaires in written form or in SL. If the SL option was selected, the materials were given by the study lead and interpreted by a professional SL interpreter. This ensured that all participants received the same content regardless of format. For participants who chose the written version, individual terms could have been translated on request. SL interpreters had received all materials in advance, along with a written briefing emphasising consistent translation in order to minimise potential bias.

3.2 Test Phases

In one of the two test phases, participants used the avatar in its fixed version. In this case, the avatar was permanently displayed above the avatar menu and anchored to the user's field of vision. In the other test phase, participants could use grab-and-drag gestures to adjust the avatar in the way that was most comfortable for them. Instructions were provided in advance of the study, ensuring all participants were familiar with the avatar operation and gestures. If necessary, the study conductor intervened during the adjustable avatar-phase to correct the use of gestures based on the instructions. To ensure comparable results, all participants interacted with the avatar while standing. The SL interpreter was present as a potential dialogue partner, serving as a point of spatial reference for the adjustments.

Each test phase included three pre-recorded animated signed sentences that participants were to select and start in the avatar menu. For one sentence (test condition I) per test phase, participants were asked to imagine themselves in a given situation, enhancing the evaluation of comprehensibility. They were given a scenario in which, for example, they were at a train station waiting for a train. The more detailed information allowed them to assume an expectation of certain contents of the subsequent sentence. In order to maintain the focus on the evaluation of comprehensibility, the two other sentences (test condition II) originated from the same superordinate scenario, but no further information on the expected content of the sentence was provided. Furthermore, these sentences were not related to each other in terms of content.

	Fixed Avatar (F)	**Adjustable Avatar (A)**
Situational comprehensibility (I)	"Unfortunately, the train is delayed by 15 minutes. It is expected to depart at 1 pm."	"ICE 1557 from Mainz, arrival 8:41 a.m., is canceled today. This is due to a technical fault on the train."
Pure Comprehensibility (II)	"Arrival RE 77 to Cologne main station via Hanover, departure originally 3:44 a.m. No stop in Cologne today."	"Track 18a, information on ICE 1557 from Mainz, arriving at 13:20, today on track 18b. I repeat: ICE 1557 from Mainz, arrival 13:20, today on track 18b. Please do not board. Arrival RE 77 to Cologne main station via Hanover, departure 3:44 a.m."

Table 2: Assigned sentences according to test conditions

3.3 Questionnaires

To capture the subjective impression of the tested avatar version immediately after each test phase and to facilitate comparison of results between all participants, two standardised questionnaires were employed.

The User Experience Questionnaire (UEQ) was designed to provide a quick and comprehensive impression of the user experience, focusing on the subjective perception of product features and their influence on users [20]. Its short version, UEQ-S was used for evaluating two versions of the avatar in one session to reduce the length of the response process and prevent fatigue-related declines in response quality.

The use of the avatar's setting options placed additional demands on users. Therefore, recording perceived stress provided insight into overall strain and contributing factors. The NASA TLX is a questionnaire that measured this stress in six dimensions [12]: mental demand, physical demand, temporal demand, performance, frustration, and effort, each recorded on a continuous scale (0 to 100) and assessed separately in 21 gradations [13]. For the calculation of total stress, the Raw TLX (RTLX) was used, where the individual dimensions were averaged, allowing for a reduction in response time. Another application of the NASA TLX was to evaluate the individual dimensions instead of the total stress, which was also considered in the evaluation.

The results of the questionnaires aimed to indicate the perceived difference between the two avatar versions. To this end, the results for the fixed avatar were compared with those for the adjustable avatar. Furthermore, a more differentiated approach was chosen: subdividing results according to avatar versions and first and second test phase enabled an evaluation of the unbiased perception in the first test phase. In addition, this differentiation made it possible to evaluate the change in results between the two test phases according to the respective anchor (F/A). A trend could be identified from the respective mean values. Confidence intervals were also evaluated with a confidence level of 95% and a one-sided t-test was used to test the tendency of the mean difference for statistical significance.

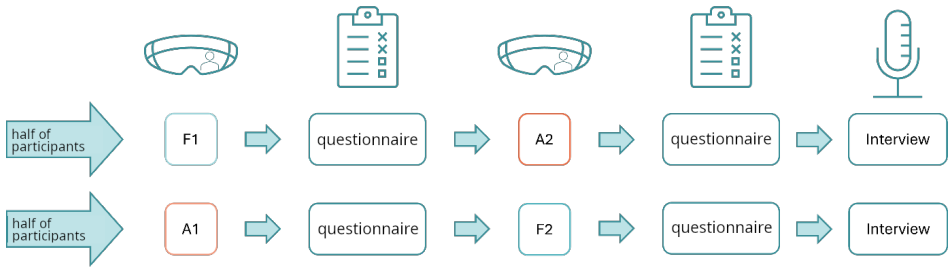


Figure 2: Procedure of the study

3.4 Comprehensibility

Although users of DGS were involved in our research team, there was not enough capacity to perform a direct analysis of the repeated signed sentence within the scope of this work. Therefore, the evaluation was done based on the transcript of the professional SL interpreters who translated the sentences repeated by the participants in DGS into spoken language. For the animated sentences of the SL avatar, there was both a transcript of the spoken language and a transcript of the SL in gloss form. There was no direct word-to-sign translation between spoken language and SL. As a result, the sentences were segmented into the smallest possible common meaning-bearing unit. This was, for example, a time, a place or a direction. The equivalent glosses and words in spoken language were assigned to this unit. The assessment itself was carried out using a performance measure in both binary and differentiated form. The comprehensibility of a unit was assigned a value. In the binary evaluation, a correctly reproduced meaning was given the value 1, otherwise the unit was given the value 0. The differentiated form was an approach to give SL a higher significance. As individual units may have consisted of multiple glosses, the following evaluation was used in this method:

Score	Criterion
0	Gloss not repeated or gloss recognised but no meaning could be assigned to it
0.5	Gloss repeated and several possible meanings were assigned, the correct meaning is below
1	Gloss correctly understood

Table 3: Evaluation criteria for comprehensibility

If a unit consisted of several glosses describing the meaning of the unit, the score was calculated for each gloss and then averaged. Due to limited resources in DGS expertise, we could not fully validate the quantification, which is why a statistical evaluation of the quantitative results was not carried out. The values represented a guideline for the test conditions under which comprehensibility was better or worse. The binary evaluation served as a control mechanism. In the further evaluation, all scores were averaged and thus resulted in an average assessment of comprehensibility. Averaging without weighting the individual units was intended to enable a neutral evaluation that avoided bias given our limited resources regarding DGS expertise. Neither communicability nor comprehension was evaluated. Therefore, a statement as to whether the units that are relevant for capturing the content of the sentence were understood is of no further significance for this work.

3.5 Interview

To gather subjective, qualitative feedback, an interview was conducted. Its aim was to understand the users' perceptions and explore the causes. In addition, this method could be used to identify approaches for further development and research. The interview was structured by three guiding questions, which could then be individually adapted to the answers in greater depth. This method focused on the *user experience* of the settings options and the *acceptability* of an avatar with settings options. At the same time, anomalies from the observation of the interaction or from the answers to the questionnaires were addressed. Finally, the participants had the opportunity to give free feedback.

The qualitative feedback consisted of observations from the session itself and the analysis of the screen recordings as well as participants' interview responses. During the evaluation process, the feedback was first clustered thematically. The aim was to examine the various areas of interaction in the avatar application and to understand the quality of use. In a second step, the importance for the application was evaluated based on findings from previous research and the frequency of occurrence. The results of the questionnaires were then combined with the qualitative feedback.

3.6 Participants

Nine individuals participated in the study, all of whom were users of SL and affiliated with the Centre for Deaf Culture and Visual Communication Berlin/Brandenburg (ZFK e.V.)[1], where the study was conducted. Four participants were aged 20-30, four were 40-50, and age was unreported for one participant. Seven participants (88.9%) were deaf; two of them used technical aids, and four regularly relied on SL interpreters. One participant was hearing and worked as a SL interpreter. Eight participants completed the study in full. One participant was unable to start the animation in the first test phase, leading to the exclusion of their quantitative data, though their qualitative feedback was retained. All participants gave a written informed consent prior to the study according to the GDPR requirements, following an approval by the DFKI ethics committee.

[1] https://zfk-bb.de/

4 Results

4.1 Comprehensibility

The evaluation of comprehensibility in binary and differentiated form came to a qualitatively similar result, only the individual values differ slightly. It did not suggest that the setting options make a difference to comprehensibility. On average, less than 50% of the individual sentence parts were understood. The situational comprehensibility of the fixed avatar was slightly higher than the average comprehensibility of all other sentences.

The transcript of the SL interpreters generally indicates that it was mainly fragments in the form of individual signs or phrases that the users managed to repeate, rarely complete sentences. In some cases, nothing was understood at all and the subjective perception that the SL avatar was difficult to understand was reported several times.

When analyzing the individual sentence components, it was noticeable that certain parts in particular tended to be more comprehensible on average. For example, times and time information were understood quite well and locations were also mostly understandable – with the exception of "Mainz". In the case of locations, however, the lack of a mouthing was particularly evident with "Hannover", since the manual sign for "Hanover" is identical to that for "blue" without a mouthing. In this case, the mouthings would have been essential and correct comprehension could only be possible if sentence context permits. This was also observed with other manual signs such as the gloss "PLAN", which was often understood as "technical". There were also still gaps in the comprehensibility of times and time units: In some cases, users only understood the hour component of the signed times, or parts of the signed numbers, such as "3" instead of "13". Such examples are obvious when the hand shapes show similarities.

The comprehensibility progression in relation to the chronology of the sentence shows that the units at the beginning were particularly difficult to understand. This effect is particularly evident with the adjustable avatar. In short sentences, it also happened that a sentence was completely missed.

When differentiating the users based on whether they actually managed to apply the adjustments, comprehensibility tends to be slightly better on average for those who had actually adjusted the avatar. Nevertheless, among them, no increase in comprehensibility from the fixed to the adjustable avatar version could be observed. The comprehensibility of this group remains largely constant. Among those who did not change the avatar, the average comprehensibility of the sentences varied more strongly overall. On average, comprehensibility is slightly lower for the adjustable version than for the fixed avatar version. However, this is more likely due to aspects of individual comprehensibility or the difficulty level of individual sentences, as otherwise the average comprehensibility for both avatar versions should not have changed for this group.

No clear correlation was found between the existing sign language levels and comprehensibility. However, this observation cannot be generalised, as all participants had a high level of proficiency in DGS. An evaluation of the average comprehensibility per participant over the course of the session showed no learning effects that could have influenced the results of comprehensibility.

Figure 3: Average comprehensibility after actual adjustment of the avatar. Values of the comprehensibility of the individual sentences after using the setting options

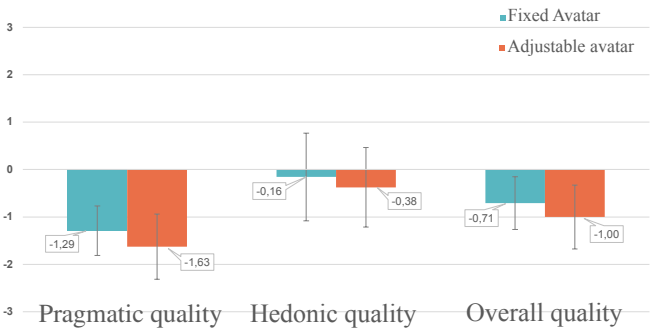


Figure 4: UEQ-S, fixed vs. adjustable avatar mean values

4.2 Evaluation of questionnaires

4.2.1 UEQ-S. Both avatar versions received an overall negative user experience rating (Figure 4). The fixed avatar was rated slightly better than the adjustable avatar, although the difference is not statistically significant. The hedonic quality of the fixed avatar was rated higher than its pragmatic quality.

When breaking down results by which version of the avatar was seen first (F1, A1) or second (F2, A2), the adjustable avatar achieved slightly more positive results than the fixed avatar (Figure 5). The hedonic quality of the adjustable avatar (A1) was rated significantly better than its pragmatic quality.

The consequent ratings show that the second test phase received lower ratings, regardless of the order in which the avatars were tested. The participants who tested the adjustable avatar first and then the fixed avatar rated the UX of the fixed avatar slightly, but not significantly worse. In contrast, the participants who first tested the fixed and then the adjustable avatar, rated the UX of the adjustable avatar significantly worse.

The results also show that if the users had actually adjusted the avatar, the user experience of the adjusted avatar was rated worse. In this subgroup, the perception of pragmatic quality was further apart between the two avatar versions than in hedonic quality.

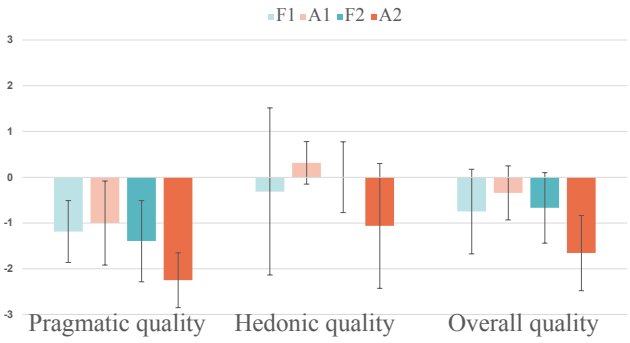


Figure 5: UEQ-S, both phases, mean value comparison

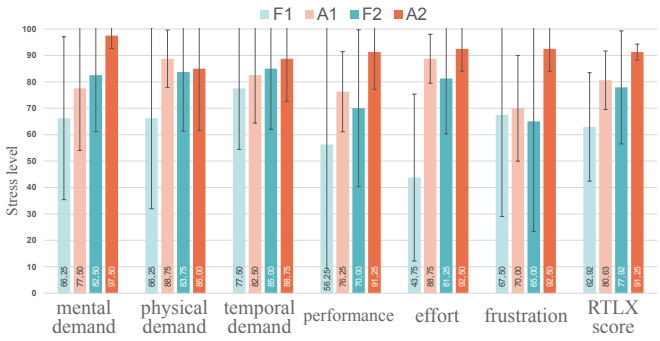


Figure 6: RTLX, differentiated mean value comparison

4.2.2 RTLX. The generally lower UX for the adjustable avatar is also reflected in the results for stress (Figure 6). The average stress of the adjustable avatar was rated significantly higher than that of the fixed avatar. The individual dimensions show that the biggest differences between the avatar versions were found in effort, performance and frustration. However, only effort was significantly higher for the adjustable avatar.

It is also noticeable that the confidence intervals for the fixed avatar were larger than for the adjustable version, since the distribution of answers was more scattered for the fixed version than for the adjustable version. The frequency distribution of the responses in the individual dimensions shows that participants assessed the adjustable avatar exclusively with stress values greater than 50, while ratings cover the full range of the scale for the assessment of the fixed avatar. This indicates that the assessment for the adjustable avatar was more uniform than for the fixed avatar, whereas the distribution of the values across the entire scale and the use of the various gradations indicate that the assessment was fundamentally differentiated.

The results of the participants who started with the fixed avatar show the lowest mean value, indicating the lowest perceived stress (Figure 6). Except for frustration, all other dimensions are the lowest compared to other test conditions. The effort dimension shows the most pronounced difference when comparing the first impression and the change to the adjustable avatar. When looking at the versions tested first, the adjustable avatar is associated with slightly higher perceived stress, but the difference is not significant

The analysis of the change in the perception of stress (RTLX score) between the fixed and the adjustable avatar shows a consistent picture: the lower perception of stress of the fixed avatar is independent of the order presented. Nevertheless, the adjustable avatar showed a significantly higher stress level only when the fixed avatar was tested first. If the adjustable avatar was tested first, the perceived stress decreased when switching to the fixed avatar, but not significantly. The adjustable avatar thus shows the highest average stress levels in the second test phase. In addition, frustration levels appear particularly high for the adjustable avatar in the second test phase.

The analysis of the individual dimensions shows that the mental demand and the time demand could have increased as the test phases progressed. On the other hand, the remaining dimensions reflect the possible increased stress caused by the avatar's adjustment options.

The results of the participants who had actually adjusted the avatar indicate the greatest difference between the fixed and adjustable avatar in perceived effort. Both groups rated the adjustable avatar as more strenuous to use. The frustration dimension shows a fairly large difference between the fixed and adjustable avatar for those who had actually adjusted it.

4.3 Qualitative results

The observations and interviews can be summarised in three areas with regard to usability: the interaction feedback of the application, the controllability of the avatar and the adjustment options.

4.3.1 Interaction feedback from the app. From the observation and the reactions of the participants, it can be deduced that the interaction feedback provided by the application was often not perceived or even entirely absent. When selecting and playing the sentences, many participants struggled to recognise whether they had successfully selected a sentence and repeatedly tapped the button even though the sentence was already loading. The "*Loading*" message was often overlooked, and some assumed the system had not responded, even though the selection was successful.

A similar phenomenon was observed when starting the animation: Users continued to focus on the "*Play animation*" button or the menu instead of looking at the avatar figure. Expecting the avatar to start signing immediately after selecting the button, users then checked the state of the avatar themselves. Not every attempt to tap the "*Play animation*" button worked, so it often took several attempts, which was described as "stressful". If the avatar figure and the menu were not displayed together in the field of view, the required head movements intensified this effect. Focusing on the menu for too long while waiting for feedback led to the beginnings of sentences or even entire sentences being missed.

Even when the participants adjusted the avatar, they lacked adequate feedback. Despite having received prior instructions on the gestures, the participants were often unsure at which adjustment points the avatar could be manipulated and whether the attempted gestures were correct. The uncertainty was intensified when the application of a theoretically correct gesture did not work.

4.3.2 Controllability of the avatar. The use of gestures led to difficulties not only with regard to feedback, but also in other situations. Some participants used gestures other than those defined in the instructions. Wrist rotations to adjust orientation and zoom gestures familiar from touchscreens to adjust the size were observed. When asked, some of these were described as more intuitive. In some cases, however, they were also tried as an alternative when the instructed gestures failed to work. The non-functioning of gestures seems to be a recurring problem: Overall, correct use of the gestures proved difficult, with some participants describing it as "overwhelming". The screen recordings show that users sometimes applied the gestures before their hand reached the adjustment points. Other times, the application did not work for unknown reasons or the adjustment could not be completed. This manifested itself in the avatar "getting lost" in the movement, even though the gripping gesture had not been released. It is possible that the gesture or hand recognition failed at these moments, but this cannot be evaluated from the available data. Nevertheless, it was noticeable that gestures had to be used several times – this was also reported back accordingly. Some participants also gave up trying to adjust the avatar using gestures after a certain time, even though the avatar had not yet been optimally adjusted to their preferences.

4.3.3 Adjustment options. When positioning the avatar, the participants grabbed it and moved it to position it within arm's reach. While the avatar is displayed at a reduced size as long as it is gripped, it "stands" on the floor once released. This natural behavior sometimes led to a significant enlargement of the avatar, which was then perceived as a combination of "too large" and "too close". In order to place the avatar in a more distant position, users would have had to move themselves and then step back after positioning the avatar. However, no user recognised and applied this.

4.3.4 Quality of the presentation. Although the participants were informed about limitations such as the lack of mouthings, as the quality of the SL avatar was not the focus of the evaluation, participants still explicitly commented on this issue. The consistent facial expressions were described as unpleasant and severely impaired the perception of the avatar. This confirms and highlights the importance of the quality of the avatar and the animation. Its rigid appearance and lack of individual style were criticised as well. During the test phases, an unnatural conversational posture was observed in some participants. In some cases, they leaned forward slightly when trying to understand the sentences or squinted their eyes, which did not occur during communication with the interpreter. Three participants also identified the avatar's legs as unnecessary and would have preferred to see only the signing space.

4.4 Expressed need for the adjustment features

When asked about whether the adjustment features are needed – assuming they were easy to use –, 56% preferred the adjustment features, 22% were undecided and 22% preferred a non-adjustable avatar. However, some participants noted that it was difficult to answer the question based on the present avatar due to its usability problems, as they had to refer only to their imagination.

5 Discussion and Conclusion

This study examined the comprehensibility, user experience, and acceptability of adding adjustment options on an SL avatar operating on Microsoft HoloLens 2. Although the majority of the users expressed a preference for the adjustment features, amidst a lot of technical problems and missing features, no improvements were observed in comprehensibility and user experience.

The experiments indicated low comprehensibility irrelevant of the adjustment features, mostly due to lack of mouthings and facial expressions, and difficulty of distinction between similar hand shapes. The adjustable version did not achieve any significant difference in comprehensibility as compared to the fixed version. It is unclear if this is because the adjustment features are not a requirement for basic understanding, or that the technical limitations were so strong that they overshadowed every other improvement. Suggestions for improvements included better interaction feedback, animated facial features, and a replay function.

For user experience, the adjustable avatar did not perform better overall than the fixed one. Challenges with gesture interaction, unclear feedback, and high effort contributed to lower UX ratings. Pragmatic quality was generally rated lower than hedonic quality, indicating that users found the system more emotionally or aesthetically pleasing than functionally useful. Stress levels were higher for the adjustable avatar, reflecting increased effort, performance and frustration levels. However, the adjustable avatar shows significantly higher stress levels only when the fixed avatar was tested first. UX tends to decline as the session progressed, likely due to repeated usability issues. The concern about whether the adjustment gestures are intuitive, and how easily users could get familiar with them, was raised. Recommendations included visual aids, tutorials, and cropping the avatar to the signing space.

In terms of acceptability, the concept of the adjustment options was preferred by the majority, but its practical value is strongly linked to usability. Poor usability leads to frustration, thereby reducing acceptance. While some concerns were raised regarding accessibility for less tech-savvy individuals, the adjustment options were seen as valuable if they were easy to use. More critically, the general acceptability of SL avatars depends heavily on animation quality and social perception—particularly concerns about avatars replacing human interpreters.

Due to the high cost of interpretation and communication barriers leading to a slow experiment speed (one hour per participant), the study had to be confined to a small sample size. This is its main limitation, as it reduces the statistical significance and affects the reliability of the quantitative results. Additionally, while necessary, interpreter involvement introduces variability that cannot be fully controlled. Lastly, the sentence material could not be customised for evaluation purposes due to system constraints.

The study confirms that SL avatar development is promising, especially with personalisation features, but highlights that usability and animation quality are key. Adjustment options alone do not guarantee improved user experience or acceptability. Importantly, comprehensibility should not rely on adjustability – avatars must be understandable by default. Future development should prioritise facial animation, better interaction design, and participatory development with deaf communities.

Acknowledgments

The research reported in this paper was primarily conducted as a BSc thesis at the Technical University of Berlin in co-operation with the German Research Center for Artificial Intelligence (DFKI) supported by BMBF (German Federal Ministry of Education and Research) via the project SocialWear (grant no. 01IW20002) and by the European Union via the project SignReality, as part of financial support to third parties by the UTTER project (Horizon Europe, GA: 101070631).

References

[1] P W Aditama, P S U Putra, I M M Yusa, and I N T A Putra. 2021. Designing Augmented Reality Sibi Sign Language as a Learning Media. *Journal of Physics: Conference Series* 1810, 1 (March 2021), 012038. doi:10.1088/1742-6596/1810/1/012038

[2] B. Alexandre, E. Reynaud, F. Osiurak, and J. Navarro. 2018. Acceptance and acceptability criteria: a literature review. *Cognition, Technology & Work* 20, 2 (2018), 165–177. doi:10.1007/s10111-018-0459-1

[3] M. Aziz and A. Othman. 2023. Evolution and Trends in Sign Language Avatar Systems: Un-veiling a 40-Year Journey via Systematic Review. *Multimodal Technologies and Interaction* 7, 10 (2023), 1–33. doi:10.3390/mti7100097

[4] L. Bernhard, F. Nunnari, A. Unger, J. Bauerdiek, C. Dold, M. Hauck, A. Stricker, T. Baur, A. Heimerl, E. André, M. Reinecker, C. España-Bonet, Y. Hamidullah, S. Busemann, P. Gebhard, C. Jäger, S. Wecker, Y. Kossel, H. Müller, and D. Wallach. 2022. Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans. In *Proceedings of the 15th International Conference on Pervasive Technologies Related to Assistive Environments*. Association for Computing Machinery, 260–268. doi:10.1145/3529190.3529202

[5] bgsd. 2020. 3rd KHV_statement BGSD. https://bgsd.de/de/verband/downloadbereich.html?file=files/documents/hinweise/3.%20KHV_Stellungnahme%20BGSD_final2.pdf&cid=1062

[6] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, J. P. Bigham, S. Azenkot, and S. K. Kane (Eds.). ACM, 16–31. doi:10.1145/3308561.3353774

[7] Michael Cabanillas-Carbonell, Piero Cusi-Ruiz, Daniela Prudencio-Galvez, and José Luis Herrera Salazar. 2022. Mobile Application with Augmented Reality to Improve the Process of Learning Sign Language. *16* (June 2022). doi:10.3991/ijim.v16i11.29717

[8] K. Crowe, M. Marschark, and S. McLeod. 2019. Measuring intelligibility in signed languages. *Clinical Linguistics & Phonetics* 33, 10–11 (2019), 991–1008. doi:10.1080/02699206.2019.1600169

[9] F. D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319. doi:10.2307/249008

[10] J. M. de Martino, I. R. Silva, C. Z. Bolognini, P. D. P. Costa, K. M. O. Kumada, L. C. Coradine, P. H. S. Da Brito, W. M. do Amaral, Â. B. Benetti, E. T. Poeta, L. M. G. Angare, C. M. Ferreira, and D. F. de Conti. 2017. Signing avatars: making education more inclusive. *Universal Access in the Information Society* 16, 3 (2017), 793–808. doi:10.1007/s10209-016-0504-x

[11] N. Fox, B. Woll, and K. Cormier. 2023. Best practices for sign language technology research. *Universal Access in the Information Society* (2023), 1–9. doi:10.1007/s10209-023-01039-1

[12] S. G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. 904–908. doi:10.1177/154193120605000909

[13] Human Performance Research Group, NASA Ames Research Center. [n. d.]. Nasa Task Load Index (TLX): Paper and Pencil Package.

[14] N. K. Kahlon and W. Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society* 22, 1 (2023), 1–35. doi:10.1007/s10209-021-00823-1

[15] Y. Kim, H. Kim, and Y. O. Kim. 2017. Virtual Reality and Augmented Reality in Plastic Surgery: A Review. *Archives of Plastic Surgery* 44, 3 (2017), 179–187. doi:10.5999/aps.2017.44.3.179

[16] M. Kipp, Q. Nguyen, A. Heloir, and S. Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, K. F. McCoy (Ed.). ACM, 107–114. doi:10.1145/2049536.2049557

[17] Ines Kožuh, Simon Hauptman, Primož Kosec, and Matjaž Debevc. 2015. Assessing the Efficiency of Using Augmented Reality for Learning Sign Language. In *Universal Access in Human-Computer Interaction. Access to Interaction*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 404–415. doi:10.1007/978-3-319-20681-3_38

[18] V. Krausneker and S. Schügerl. 2022. Avatars for Sign Languages: Best Practice from the Perspective of Deaf Users. In *Assistive Technology, Accessibility and (e)Inclusion*, A. Petz, E.-J. Hoogerwerf, and K. Mavrou (Eds.). Association ICCHP, 156–164. doi:10.35011/ICCHP-AAATE22-P1-21

[19] I. Lacerda, H. Nicolau, and L. Coheur. 2023. Towards Realistic Sign Language Animations. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, B. Lugrin, M. Latoschik, S. von Mammen, S. Kopp, F. Pécune, and C. Pelachaud (Eds.). Association for Computing Machinery, 1–4. doi:10.1145/3570945.3607354

[20] B. Laugwitz, T. Held, and M. Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work (Lecture Notes in Computer Science, Vol. 5298)*, A. Holzinger (Ed.). Springer, 63–76. doi:10.1007/978-3-540-89350-9_6

[21] Qiwei Liang, Yikeng Chen, Wenbiao Li, Minghao Lai, Wenjian Ni, and Hong Qiu. 2024. iKnowiSee: AR Glasses with Language Learning Translation System and Identity Recognition System Built Based on Large Pre-trained Models of Language and Vision and Internet of Things Technology. In *Intelligent Networked Things*, Lin Zhang, Wensheng Yu, Quan Wang, Yuanjun Laili, and Yongkui Liu (Eds.). Springer Nature, Singapore, 12–24. doi:10.1007/978-981-97-3948-6_2

[22] Le Luo, Dongdong Weng, Guo Songrui, Jie Hao, and Ziqi Tu. 2022. Avatar Interpreter: Improving Classroom Experiences for Deaf and Hard-of-Hearing People Based on Augmented Reality. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–5. doi:10.1145/3491101.3519799

[23] V. López-Ludeña, R. San-Segundo, J. M. Montero, R. Córdoba, J. Ferreiros, and J. M. Pardo. 2012. Automatic categorization for improving Spanish into Spanish Sign Language machine translation. *Computer Speech & Language* 26, 3 (2012), 149–167. doi:10.1016/j.csl.2011.09.003

[24] M. Mori, K. F. MacDorman, and V. Schwind. 2019. The uncanny valley. Translation from the Japanese. Zenodo. doi:10.5281/zenodo.3226987

[25] S. Möller. 2017. *Quality Engineering: Quality of communication technology systems* (2nd ed.). Springer. doi:10.1007/978-3-662-56046-4

[26] M. Müller, Z. Jiang, A. Moryossef, A. Rios, and S. Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 682–693. doi:10.18653/v1/2023.acl-short.60

[27] Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. 2021. Automatic Generation of a 3D Sign Language Avatar on AR Glasses given 2D Videos of Human Signers. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Association for Machine Translation in the Americas, Virtual, 71–81. https://aclanthology.org/2021.mtsummit-at4ssl.8

[28] Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. 2021. Evaluating the Translation of Speech to Virtually-Performed Sign Language on AR Glasses. In *Proceedings of the Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)*. 141–144. doi:10.1109/QoMEX51781.2021.9465430

[29] J. Nielsen. 1993. *Usability Engineering*. Merchant.

[30] A. Nolte, B. Gleißl, J. Heckmann, D. Wallach, and N. Jochems. 2023. "I Want To Be Able To Change The Speed And Size Of The Avatar": Assessing User Requirements For Animated Sign Language Translation Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, A. Schmidt (Ed.). Association for Computing Machinery, 1–7. doi:10.1145/3544549.3585675

[31] Fabrizio Nunnari, Eleftherios Avramidis, Vemburaj Yadav, Alain Pagani, Yasser Hamidullah, Sepideh Mollanorozy, Cristina España-Bonet, Emil Woop, and Patrick Gebhard. 2023. Towards Incorporating 3d Space-Awareness into an Augmented Reality Sign Language Interpreter. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 1–5. https://ieeexplore.ieee.org/abstract/document/10193194/

[32] F. Nunnari, J. Bauerdiek, L. Bernhard, C. España-Bonet, C. Jäger, A. Unger, K. Waldow, S. Wecker, E. Andre, S. Busemann, C. Dold, A. Fuhrmann, P. Gebhard, Y. Hamidullah, M. Hauck, Y. Kossel, M. Misiak, D. Wallach, and A. Stricker. 2021. AVASAG: A German Sign Language Translation System for Public Services. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. 43–48. https://aclanthology.org/2021.mtsummit-at4ssl.5/

[33] A. Othman, A. Dhouib, H. Chalghoumi, O. El Ghoul, and A. Al-Mutawaa. 2024. The Acceptance of Culturally Adapted Signing Avatars Among Deaf and Hard-of-Hearing Individuals. *IEEE Access* 12 (2024), 78624–78640. doi:10.1109/ACCESS.2024.3407128

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311. doi:10.3115/1073083.1073135

[35] F. Picron, D. van Landuyt, R. Omardeen, E. Efthimiou, R. Wolfe, S.-E. Fotinea, T. Goulas, C. Tismer, M. Kopf, and T. Hanke. 2024. The EASIER Mobile Application and Avatar End-User Evaluation Methodology. In *Proceedings of*

the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources. 276–281. https://aclanthology.org/2024.signlang-1.31/

[36] Lorna Quandt. 2020. Teaching ASL Signs Using Signing Avatars and Immersive Learning in Virtual Reality. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3373625.3418042

[37] L. C. Quandt, A. Willis, M. Schwenk, K. Weeks, and R. Ferster. 2022. Attitudes Toward Signing Avatars Vary Depending on Hearing Status, Age of Signed Language Acquisition, and Avatar Type. *Frontiers in Psychology* 13 (2022), 1–14. doi:10.3389/fpsyg.2022.730917

[38] R. G. Smith and B. Nolan. 2016. Emotional facial expressions in synthesized sign language avatars: a manual evaluation. *Universal Access in the Information Society* 15, 4 (2016), 567–576. doi:10.1007/s10209-015-0410-7

[39] Noor-Un-Nissah B N Soogund and Minnu Helen Joseph. 2019. SignAR: A Sign Language Translator Application With Augmented Reality. *Journal of Applied Technology and Innovation* 3, 1 (2019), 33–37.

[40] I. Wechsung and K. de Moor. 2014. Quality of Experience Versus User Experience. In *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake (Eds.). Springer, 35–54. doi:10.1007/978-3-319-02681-7_3

[41] R. Wolfe, A. Braffort, E. Efthimiou, E. Fotinea, T. Hanke, and D. Shterionov. 2023. Special issue on sign language translation and avatar technology. *Universal Access in the Information Society* (2023), 1–3. doi:10.1007/s10209-023-01014-w

[42] R. Wolfe, J. C. McDonald, T. Hanke, S. Ebling, D. van Landuyt, F. Picron, V. Krausneker, E. Efthimiou, E. Fotinea, and A. Braffort. 2022. Sign Language Avatars: A Question of Representation. *Information* 13, 4 (2022), 206. doi:10.3390/info13040206

[43] World Federation of the Deaf. 2021. Our Work - WFD. https://wfdeaf.org/our-work/ March 19.

[44] F.-C. Yang, C. Mousas, and N. Adamo. 2022. Holographic sign language avatar interpreter: A user interaction study in a mixed reality classroom. *Computer Animation and Virtual Worlds* 33, 3–4 (2022), e2082. doi:10.1002/cav.2082

[45] Fu-Chia Yang, Christos Mousas, and Nicoletta Adamo. 2022. Holographic Sign Language Avatar Interpreter: A User Interaction Study in a Mixed Reality Classroom. *Computer Animation and Virtual Worlds* 33, 3-4 (2022), e2082. doi:10.1002/cav.2082