





The Efficiency of Augmented Pointing with and Without Speech in a Collaborative Virtual Environment

Oliver Herbert^(✉)  and Lisa-Marie Krause 

Department of Psychology, University of Würzburg, Würzburg, Germany
oliver.herbert@uni-wuerzburg.de

Abstract. Pointing is a ubiquitous part of human communication. However, pointing gestures to distal referents are often misunderstood systematically, which may limit the usefulness of pointing. We examined pointing-based communication in a collaborative virtual environment (CVE) to address three questions. First, we wanted to evaluate the potential of apparently natural but technically augmented pointing in CVEs, such as presenting a warped pointer for increased legibility or the ability to assume the pointer's perspective. Second, we wanted to test whether technical improvements in pointing accuracy also facilitate communication if pointing is accompanied by speech. Third, we wanted to check whether pointing accuracy is correlated with the efficiency of communication involving pointing and speech. An experiment revealed that pointing-based communication has considerable potential to be enhanced in CVEs, albeit the specific augmentation procedure we employed did not improve pointing-based communication. Importantly, improvements in pointing accuracy also facilitated communication when speech was allowed. Thereby, speech reduced but could not rule out misunderstandings. Finally, even a small gain in pointing accuracy allowed participants to agree on a referent faster. In summary, the experiment suggests that augmented pointing may considerably improve interactions in CVEs. Moreover, speech cannot fully compensate misunderstandings of pointing gestures and relatively small differences in pointing accuracy affect the efficiency of communication of speech is allowed.

Keywords: Pointing · Collaborative virtual environment · Speech · Multimodal reference · Deixis

1 Introduction

When people interact, they frequently use pointing gestures to refer to locations or objects in the environment. Pointing gestures appear especially helpful, when verbal descriptions are unlikely to quickly guide the attention of others to the referent. This

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-35741-1_37.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
V. G. Duffy (Ed.): HCII 2023, LNCS 14028, pp. 1–15, 2023.
https://doi.org/10.1007/978-3-031-35741-1_37

may be the case when the object is not salient (e.g., a wild animal hidden in a distant tree line) or when the vocabulary to describe the object is lacking (e.g., shopping pastries in a foreign country). Thus, it comes as no surprise that attempts have been made to facilitate pointing in collaborative virtual environments (CVE) [11, 16, 18, 19]. In the present paper, we want to examine the effectiveness and efficiency of pointing-based communication with and without speech in CVEs.

1.1 Pointing Accuracy

As helpful as pointing may be, pointing does not always guarantee accurate reference in situations, in which a relatively high degree of pointing accuracy would be required while it is not possible to touch the referent. This has mainly two reasons, which result from the necessity to extrapolate from the gesture toward potential referents. First, the process introduces unsystematic errors. That is, the same gesture may be interpreted in different ways in different situations. Second, the interpretations of pointing gestures are typically biased [3] as pointing production and interpretation follow different geometric rules. A pointer typically puts their index finger on the line between their eyes and the referent [6, 17]. Consequently, the eyes of the pointer, index finger, and referent form a line. By contrast, observers typically use a combination of two different visual cues [8]. When beside the pointer, observers typically extrapolate from the arm and index finger, which results in upward biased interpretations.¹ The more observers approach the perspective of the pointer, the more they rely on the vector defined by their (the observers’) eyes and the pointer’s index finger. This mode of extrapolation typically introduces sideward biases. Additional vertical biases may result when the eye heights of pointer and observer differ. Besides these perspective-induced biases, other biases exist such as non-linearities in human extrapolation [6].

These systematic biases usually contribute considerably to the errors that emerge during pointing-based communication. Hence, different attempts have been made to reduce them. For example, Herbort and Kunde used instructions to improve the legibility of pointing gestures or observers’ interpretations thereof [7]. An ingenious method for compensating biases in pointing perception in collaborative virtual environments has been suggested by Sousa and colleagues [16]. They presented a distorted or “warped” version of a pointer in the virtual environment of the observer. The virtual pointer was transformed in such a way, that pointing gestures could be understood better by a naïve observer. This approach has the advantage that the legibility of pointing gestures is improved but the gestures still appear natural – which may be helpful in scenarios that aim for realism. The method has since then been improved to counteract horizontal and vertical biases [11]. By contrast, other methods such as laser pointers, spotlights, and very long arms [19] may be more effective means of reference but lack realism. Thus, one central question of the present paper is to further examine the potential of the presentation of warped virtual pointers.

¹ In the current paper, we examine misunderstandings or errors between pointers and observers without considering the individual contributions of both interlocutors. If we use terms such as “biased interpretations”, we refer to the mismatch between the pointer’s referent and the observer’s guess thereof without attributing the misunderstanding to either of the interlocutors.

1.2 Pointing and Speech

In many studies referenced above, pointing was examined in isolation [3, 6–8, 17]. By contrast, pointing is usually accompanied by speech in natural situations. Pointers thus could provide additional verbal descriptions that would allow observers to discern whether they correctly identified the referent. Hence, speech has the potential to eliminate the inevitable inaccuracies of pointing gestures. Moreover, it has been suggested that pointing gestures refer mostly to larger area of space [4] rather than specific locations or small areas [10]. In this case, fine-grained improvement of pointing legibility may be shrouded by verbally conveyed information.

This raises the question whether speech compensate the inaccuracies of pointing and if so, to which extent more accurate pointing nevertheless facilitates communication. With respect to the first question, previous research indicates that pointing per se is a helpful tool for reference, which, however, may be partially compensated by verbal descriptions [9]. The expected accuracy of pointing as such also plays a role. For example, when the distance between interlocutors and an array of potential referents is increased, also more words, especially location descriptions, are used to single out the referent [2, 12]. At the same time, the frequency of pointing gestures decreases.

The second question is more difficult to address because the accuracy of pointing is typically difficult to manipulate in isolation. In previous experiments, changes in pointing accuracy resulted from other changes in the situation (e.g., distance to referents [2]), which in turn affected the types of descriptions used, the prevalence of pointing as such, and the expectations of participants. This makes it difficult to pin-point the effect of pointing accuracy on the effectiveness and efficiency of pointing-based communication. The warping technique allows a unique opportunity to examine the isolated effect of pointing accuracy, because it allows to manipulate pointing accuracy without any apparent changes in the environment. Thus, neither the pointer nor the observer may be aware or even able to detect this type of manipulation.

1.3 The Current Experiment

In the current experiment, we examined pointing-based communication in a CVE. The CVE was modelled after a planetarium, in which one person (pointer) was asked to help another person (observer) find various predetermined planets in an array of similar planets. Our aims were threefold.

First, we wanted to evaluate augmented pointing methods that reduce biases in pointing perception. We evaluated four pointing modes in conditions, in which participants were asked not to speak. In a *naïve mode*, the pointer's body postures were presented veridically in the observer's virtual environment. In a *warping mode*, the pointer's posture was transformed to increase pointing legibility [11, 16]. Both modes were indissociable from the viewpoint of the participants. Next, we used a *pointer perspective mode* in which the observer could assume the perspective of the pointer. It has been shown that this virtually eliminates systematic misunderstandings and reduces unsystematic variability [5]. This mode thus can serve as an upper-bound of the performance that could be achieved with elaborate warping techniques and hence with a natural appearing virtual representation of the pointer. Finally, as a comparison, we included a *laser pointer*

mode in which referents can be indicated unambiguously. We expected minimal errors in the laser pointer condition, smaller biases and total errors in the warping and pointer perspective conditions, and larger biases and total errors in the naïve condition.

Second, we wanted to check whether and to which extent the reduction in biases by augmented pointing also affects the accuracy and speed of communication when accompanied by speech. To this end, the different pointing modes were combined with the possibility to speak. If speech fully compensates for misunderstandings, few if any misunderstandings should occur in the speech conditions. If speech is insufficient to rule out misunderstandings, we expect that speech reduces but does not eliminate misunderstandings. In this case, misunderstandings should be largest in the naïve pointing mode, medium in the warping and pointer perspective mode, and smallest in the laser pointer mode. If the accuracy of pointing furthermore affects the efficiency of the communication, we would expect that participants are fastest in the laser pointer mode, slower in the warping and pointer perspective mode, and slowest in the naïve mode. Moreover, the apparent accuracy of pointing is expected to affect how participants communicate verbally.

Third, we were interested in the relationship between overall pointing accuracy and the efficiency in establishing a joint focus of attention based on pointing and speech. To this end, we correlated pointing accuracy in conditions without speech with the speed in the respective conditions across participants.

2 Methods

2.1 Participants

Forty-eight persons (13 male, 34 female, 41 right-handed, 6 left-handed, mean age 28 years range 20 to 67, one participant did not disclose gender, handedness, and age) or 24 pointer-observer dyads from the Würzburg area participated in exchange for course credit or money. The experiment was in accordance with the standards of the ethics committee of the Department of Psychology of the University of Würzburg.

2.2 Stimuli and Apparatus

Both members of a dyad participated in the same room. They saw visual representations of each other in VR but could directly hear the other participant's voice. Both participants were seated on chairs facing the same direction with a lateral distance of 100 cm. The observer always sat to the right of the pointer and in front of a small desk. Both participants were immersed in a virtual environment with a HTC Vive Pro (observer) or HTC Vive Pro Eye (pointer) head mounted display (HMD). Both HMDs have identical displays. The Vive Pro Eye is equipped with an additional eye tracker, which was used to visualize the eye-movements of the pointer. The pointer wielded two HTC Vive wireless Controllers. Additionally, HTC Vive Trackers were attached to the pointer's upper arms. The two PCs controlling the experiment for the pointer and observer were connected via LAN. The experiment was created with the Unity 3D engine.

The VR environment consisted of a virtual planetarium, in which planets were presented on an imaginary sphere with radius 300 cm (Fig. 1, Supplemental Video 1). Planets were arranged in a grid of 37 columns (azimuths of -90° to 90° in 5° steps) and 16 rows (elevations of -15° to 60° in 5° steps). Planets were random combinations of the features diameter (10.0 cm, 12.5 cm, 15.0 cm), texture (Venus, Mars, Jupiter, Saturn, Neptune, Pluto, textures from nasa3d.arc.nasa.gov/images), and moons (none, small moon with texture of Phobos, large moon with texture of Phobos, large moon with texture of Charon, two moons with textures of Phobos and Charon). Planets were presented at random orientations. Hence, planets with identical size, texture, and moon configuration differed with respect to the part of the texture facing the participant, the relative position of the moons, and adjacent planets. The planets were newly generated at the beginning of each block but did not change within blocks. For the pointer, the target planet was indicated by a green square-shaped marker. The observer could move a similar green marker (cursor) with the mouse to select planets. The cursor moved on the planet sphere. Both participants could not see the cursor or marker of the respective other participant.

The avatars were placed on a wooden bench (160 cm x 40 cm x 51 cm) at the center of the planet sphere. The distance of the avatars corresponded to the distance of the participants in the lab. The bench was positioned on a wooden floor. The pointer's avatar was presented as highly stylized figure consisting of a head, a torso, and two arms and hands, which formed a pointing gesture. The observer's avatar only consisted of a head. The position and orientation of the HMDs was mapped on the avatars' heads. In addition, the pointer's avatar reflected the pointer's eye and lid movements. The position and orientation of the pointer's handheld controllers and trackers were mapped onto the pointer's virtual hands, wrists, elbows, and shoulders.

In the laser pointer condition, the pointer's right index finger emitted a red transparent laser ray that ended on the planet sphere. The ray was visible for pointers and observers. In the warping condition, the pointer's arms and hand were presented unchanged to the pointer but were presented rotated to the observer. The rotation was computed as follows.² First, the pointed-at position A was estimated by extrapolating the vector defined by the pointer's cyclopean eye and their right index fingertip toward the planet sphere. Next, the observer's naïve guess B was estimated by extrapolating the pointer's right index finger to the planet sphere. The angle between A, the pointer's right shoulder, and B was computed. The arm was then rotated around the shoulder by this angle, so that the positions A and the rotated position B' would be aligned from the observer's viewpoint. Note that this procedure does not use information about the exact referent except for its distance.

² Our algorithm differed from previously applied warping methods [11,16]. Unlike [16], it takes horizontal errors into account but does not consider the non-linearity of pointing extrapolation. Unlike [11], we used a parameter-free geometric model for simplicity.

2.3 Procedure

Upon providing informed consent, the participants were randomly assigned the role of pointer or observer. Both participants received the HMDs. Next, the pointer was asked to stay in a steady posture while the positions of the shoulders, elbows, wrists, and fingertip relative to the tracker position were determined with a third tracker to allow an accurate mapping of the pointer’s posture to their avatar. The experiment was split into eight blocks. At the beginning of each block, the participants were instructed orally by the experimenter. From the view of the pointer, a trial began when a planet was enclosed by the green marker. From the view of the observer, the trial began when the cursor appeared, which could be moved with the mouse. In the first 0.25 s of the trial, cursor and marker zoomed in on their respective positions to facilitate visual detection of the marker or cursor. The trial lasted until the observer had marked one planet with the cursor and pressed the left mouse button. Then marker and cursor disappeared for one second before the next trial was started.

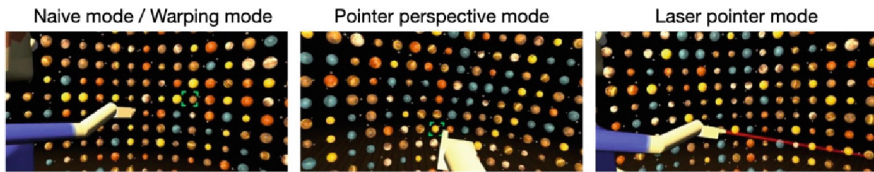


Fig. 1. The different pointing modes from the observer’s perspective.

The blocks differed with respect to the modality and the pointing mode. There were two modality conditions. In the *pointing-only* blocks, the pointer was asked to point at the referent and the observer was asked to guess the pointed at planet as good as possible. Participants were instructed not to speak. In the *pointing + speech* blocks, the pointer could point. Both participants were additionally allowed to talk to each other but were asked not to describe the planet with the help of the rows and columns of the planet grid (e.g., “third planet from the right in the fourth row from the top”) and to not refer to the planet position of previous trials.

We presented four different pointing mode conditions. In the naïve mode, the avatars of pointer and observer were presented unaltered. In the warping mode, the pointer’s right arm was presented unaltered to the pointer but rotated as described above to the observer. In the pointer perspective mode, observers could assume the viewpoint of the pointer by holding down the right mouse button. More specifically, the position of the virtual viewpoint of the observers was aligned to that of the pointer but the observer’s head rotations still controlled the direction of the observer’s virtual viewpoint. In the laser pointer mode, a laser beam protruded from the pointers finger. The laser beam was visible for pointers and observers.

The experiment was split into eight blocks. For one half of the participants, the first four blocks were *pointing-only* blocks and the remaining blocks were *pointing + speech* blocks. For the other half, this order was reversed. Pointing modes were randomly ordered within the first and second half of the experiment. Each block consisted of two

warm-up trials followed by 21 test trials. Targets were presented at azimuths of -45° , -30° , -15° , 0° , 15° , 30° , and 45° and elevations of 0° , 15° , and 30° . Each combination of azimuth and elevation was presented once. Trial order was randomized. The warm-up trials were drawn from the set of possible test trials. During the experiment, the experimenter coded the occurrence of different categories of verbal descriptions using a keyboard. The categories were features (e.g., “It’s the *small, red* planet”), locations (e.g., “It’s in the *middle*.”), deictic expressions (“It’s *this* planet.”), instructions (e.g., “You need to look left”), and off-topic themes.

2.4 Data Analysis and Reduction

For each trial, the click position of the observer’s guesses, the reaction time (defined as the interval from when the marker finished zooming in on the referent and the moment the observer clicked on a planet), and the speech content were recorded. Click positions were recorded in angular coordinates (azimuth, elevation) on the planet sphere and were rounded to steps of 5° (i.e., set to the closest planet). To measure systematic misunderstandings, horizontal and vertical errors were computed as the mean signed difference between referent and guess for azimuth and elevation, respectively. Positive values imply rightward and upward biases of the observer. To measure overall pointing accuracy, we computed the average total angle between guess and referent on the planet sphere (total error) and the percentage of clicks on the correct planet. For comparison, an error of 1° corresponds to approximately 5 cm in absolute distances. For each category of speech content, we coded whether it was used in a trial at least once. After visual inspection of the data, trials were discarded if total errors exceeded 60° (5 trials) or if reaction times were less than 1 s (1). In total, 4026 trials were entered into the analysis. Data were averaged over dyads, modalities and pointing modes. We analyzed the data with R (version 4.2.1[13]) and computed ANOVAS with the afex package (version 1.1–1[14]). Data, scripts and supplemental material are provided on <https://osf.io/97xb2/>.

3 Results

3.1 Errors and Reaction Times

The 2D-histograms of Fig. 2 give an impression of the errors made in the different conditions. Three things are noteworthy. First, except for the laser pointer mode, errors were frequent and often quite large in the pointing only condition. Second, the possibility to speak reduced but did not eliminate errors. Third, there were no errors in the laser pointer condition. Errors and reaction times were analyzed with within-participant ANOVAs with factors of pointing mode and modality, using a Greenhouse-Geisser correction if applicable. Significant results were followed up with t-tests comparing the naïve condition to the other pointing modes for each level of modality ($\alpha = .05$). Significant results are indicated by lines in Fig. 3. The full statistics are listed in Supplemental Table 1.

Horizontal Errors

Figure 3A shows systematic horizontal errors. Not surprisingly, horizontal errors were smaller when speech was allowed, $F(1,23) = 7.88$, $p = .010$, $\eta_p^2 = .26$. The pointing

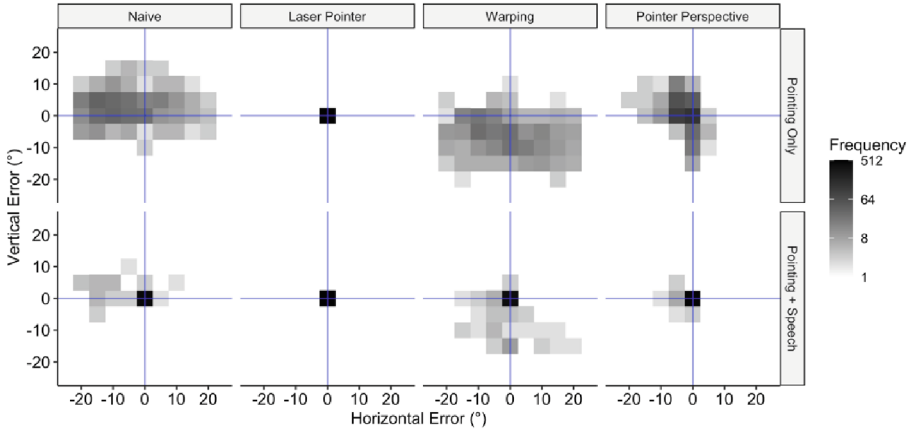


Fig. 2. 2D-Histograms of errors by pointing mode and modality. The blue lines indicate correct interpretations of pointing gestures. The color scale is logarithmic.

mode affected errors, $F(1.6,37.2) = 13.40$, $p < .001$, $\eta_p^2 = .37$. Both factors interacted, $F(1.6,37.7) = 7.69$, $p = .003$, $\eta_p^2 = .25$. Significant ($p < .05$) results of follow-up t-tests are marked as lines in Fig. 3A. Pointing gestures were interpreted as more leftward in the naïve mode compared to the other modes in the pointing only condition. In the pointing and speech condition, the effect was only preserved when comparing the naïve with the warping mode.

Vertical Errors

There was no significant main effect of modality on vertical errors, $F(1,23) = 13.97$, $p = .217$, $\eta_p^2 = .07$. However, the pointing mode affected vertical errors, $F(2.3,52.6) = 67.88$, $p < .001$, $\eta_p^2 = .75$. Pointing mode and modality interacted, $F(2.5,57.5) = 41.15$, $p < .001$, $\eta_p^2 = .64$. Pointing gestures were interpreted as more upward in the naïve mode compared to the other modes in the pointing only condition. In the pointing and speech condition, the naïve mode resulted in more upward interpretations than the laser pointer mode, warping mode, and marginally also the pointer perspective mode ($p = .083$).

Total Errors

Total errors (Fig. 3C) were considerably smaller when participants could speak, $F(1,23) = 7.95$, $p < .001$, $\eta_p^2 = .93$. In addition, the pointing mode affected total errors, $F(2.5,65.4) = 101.30$, $p < .001$, $\eta_p^2 = .82$. Pointing mode and modality interacted, $F(2.1,47.5) = 65.13$, $p < .001$, $\eta_p^2 = .74$. Observers were more accurate in the laser pointer and pointer perspective mode for both modalities. However, the warping mode resulted in more errors than the naïve mode, regardless of whether speech was allowed.

Percent Errors

In the naïve and warping mode, observers were rarely able to select the correct planet when speech was not allowed (Fig. 4D). Error rates were greatly reduced when speech was allowed, $F(1,23) = 575.39$, $p < .001$, $\eta_p^2 = .96$. The pointing mode likewise affected the percentage of errors, $F(2.1,49.1) = 185.91$, $p < .001$, $\eta_p^2 = .89$. Pointing mode

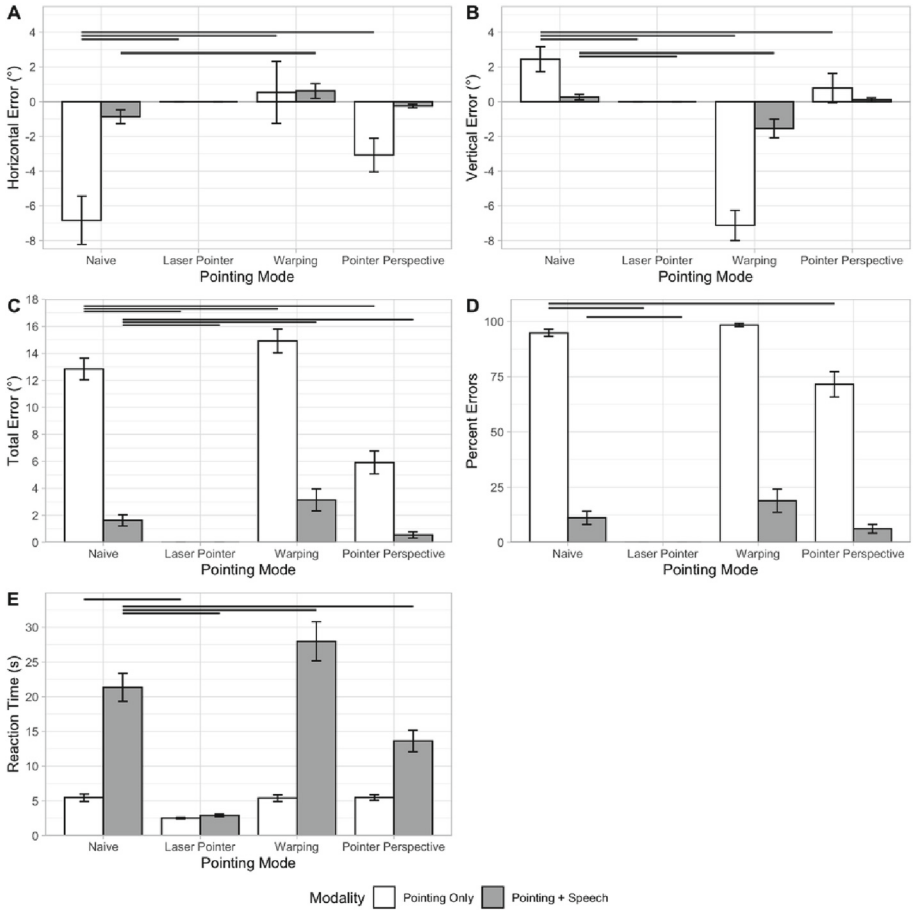


Fig. 3. Mean errors (A-D) and reaction times (E) by pointing mode and modality. No errors were made in the laser pointer mode. The horizontal lines indicate significant differences between conditions. The full statistics of the respective tests are reported as Supplemental Table 1. Positive horizontal and vertical errors are rightwards (A) and upwards (B), respectively.

and modality interacted, $F(2.1,47.9) = 99.89$, $p < .001$, $\eta_p^2 = .81$. Follow-up t-tests revealed a higher accuracy of the laser pointer mode compared to the naïve mode for both modalities. The pointer perspective mode also helped reducing errors when speech was not allowed. The remaining comparisons approached significance (all $ps \leq .062$).

Reaction Times

Figure 3E shows reaction times. Not surprisingly, reaction times were considerably shorter in the pointing-only condition, $F(1,23) = 85.41$, $p < .001$, $\eta_p^2 = .79$. Reaction times depended on the pointing mode, $F(2.4,55.1) = 53.51$, $p < .001$, $\eta_p^2 = .70$. Pointing mode and modality interacted, $F(2.3,52.6) = 45.85$, $p < .001$, $\eta_p^2 = .56$. Follow-up t-tests revealed that reaction times in the naïve mode were shorter than those in the warping

condition but longer than those in the laser pointer and pointer perspective conditions when speech was allowed. In the pointing only condition, only the laser pointer mode yielded faster reactions than the naïve mode.

3.2 Relationship Between Errors and Reaction Time

Next, we address the question how participants benefitted from pointing gestures, by comparing total errors of the pointing only condition with the reaction times in the corresponding pointing and speech condition. We computed repeated measures correlations to assess the relationship of both variables in each dyad across different pointing modes (rmcorr package, version 0.5.4 [1]). Figure 4 shows a scatter plot of both variables for each dyad and pointing mode. The analysis revealed a positive correlation between pointing accuracy and reaction times when all pointing modes were considered, $r_{rm}(71) = .78$, 95% CI [.67, .86], $p < .001$. A significant correlation was also found when only the data of the naïve and warping mode were entered into the analysis, $r_{rm}(23) = .43$, 95% CI [.04, .71], $p < .032$. In both cases, an increase in total pointing accuracy of 1° shortened reaction times by approximately 1.4 s.

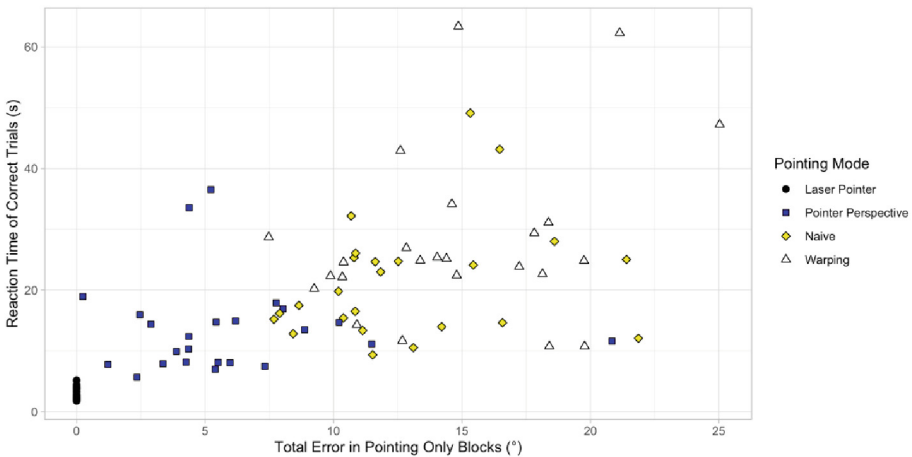


Fig. 4. Relationship between pointing accuracy and reaction time

3.3 Speech Content

Finally, we examined the speech content across pointing modes. Figure 5 shows the percentage of trials in which different types of verbal expression were used. Data was analyzed with an ANOVA with factors of pointing mode and speech content. The category “other” was not included in the analysis. Follow-up t-tests between the naïve mode and the other modes for each type of expression are provided in Supplemental Table 2 and depicted as lines in Fig. 5. The pointing mode affected the prevalence of expressions, $F(2.2, 49.8) = 89.60$, $p < .001$, $\eta_p^2 = .80$. T-tests revealed that overall, features

were described more often than locations, locations descriptions were more common than deictic expressions, and deictic expressions were more frequent than instructions, all $t(23) > 4.00$, all $ps < .001$, all $d_zs \geq 0.82$. Likewise, the different types of speech content were used with varying frequencies, $F(1.8, 41.7) = 293.45$, $p < .001$, $\eta_p^2 = .93$. Both factors interacted, $F(4.6, 105.5) = 33.35$, $p < .001$, $\eta_p^2 = .59$. The interaction can be mainly attributed to location descriptions and deictic expressions being reduced stronger than feature descriptions in the laser pointer mode compared to the other pointing modes. Follow-up t-test revealed that features, locations, deictic expressions, and instructions were less common in the laser pointer condition than the naïve condition. Additionally, more instructions were given in the warping condition compared to the naïve conditions.

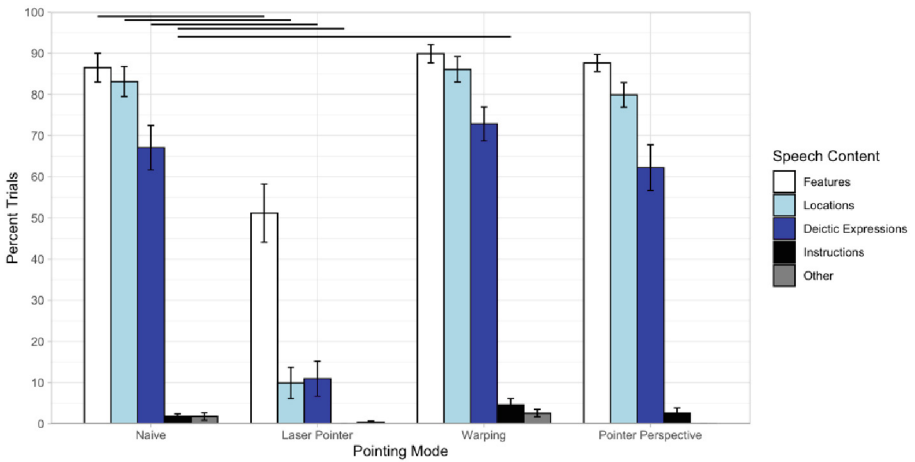


Fig. 5. Percentage of trials in which different types of speech content were used by the dyads. Lines indicate significant differences.

4 Discussion

We studied the efficiency and effectiveness of pointing-based communication with and without speech in a CVE to address three questions. Our first question was whether apparently natural pointing can be augmented to increase pointing accuracy. The answer is mixed. On the one hand, the warping mode reduced horizontal errors but increased vertical errors, resulting in an overall worse performance in our experiment. On the other hand, the presence of systematic errors per se and the decent accuracy of the pointer perspective mode, which indicates the level of accuracy that might be approximated with improved warping techniques, suggests that augmented presentations of pointing gestures may considerably facilitate interactions in CVEs.

Our second question addressed whether accuracy differences between pointing modes are leveled out by speech and whether such differences affect the efficiency when communicating with pointing and speech. Not surprisingly, it became evident that

speech allows to greatly reduce the number of errors. However, errors were still made in a substantial percentage of trials (e.g., 11% in the naïve mode) and differed between conditions. Thus, speech can only partially compensate misunderstandings of pointing gestures. Moreover, the accuracy of the different pointing modes without speech were related to the efficiency of communication in the pointing and speech conditions. For example, participants required on average about 14 s to agree on a planet in the relatively accurate pointer perspective mode but twice as long in the warping mode. The different modes of pointing also affected the content of speech. Not surprisingly, locations were rarely described if pointing was unambiguous but frequently used in the other conditions. This supports the notion that pointing as a method of reference is an integral part of otherwise verbal communication [9] and does not only supplement speech with principally redundant gestures.

Thirdly, and related to the previous point, the accuracy of pointing of a specific dyad in a specific pointing mode without speech was tightly correlated with the respective performance in the condition with speech. The analysis showed that even small increments in pointing accuracy directly facilitated verbal interactions.

4.1 Implications for Augmenting Pointing

In the present experiment, pointing interpretations were biased left-ward and up-ward. This result is consistent with the notion that pointers align eye, fingertip, and referent whereas observers extrapolate the pointer's index finger vector [6, 17]. Thus, the behavior in the naïve mode replicates previous findings in real and virtual settings [3, 5–8, 17].

We included a warping mode that was intended to counteract these biases and thus reduce the average magnitude of errors in pointing-based communication. The warping mode indeed shifted systematic errors. In the naïve mode, horizontal errors were relatively large and were considerably reduced by the warping mode. However, vertical errors were smaller than expected and resulted in an overcompensation by the warping mechanism. This led to a strong downward bias and an increment of total errors and the percentage of errors. However, given that vertical errors are typically higher than in our experiment and that the warping technique has been shown to reduce errors elsewhere [11, 16], we would argue that it is – in principle – a feasible technique. However, the parameter-free application in the present experiment may have been an oversimplified approach. First, the present data revealed some variability between systematic errors of the different dyads. Second, different realizations of CVEs and the representation of avatars may also affect biases in pointing interpretations. Hence, a calibration of warping algorithms to different users and environments may prove useful [11]. Third, using elaborate models of pointing perception, which account for other biases, might further improve performance and the generalizability of warping mechanisms [6, 8].

Pointer-observer misunderstandings are mostly rooted in the difference in viewpoints [5]. Hence, we introduced the pointer perspective mode in which both interlocutors assume the same perspective and in which systematic misunderstandings appear to be minimal. As participants rely mostly on the position of the index finger in their visual fields in this condition, visual uncertainty associated with extrapolating the finger is small in this condition [5]. Thus, this mode can be seen as an upper boundary for the accuracy that can be achieved with a sophisticated morphing method. Indeed, accuracy

was much higher in this condition than in the naïve mode and participants could agree on a planet much faster if speech was allowed. Thus, given these increments, it seems a worthwhile endeavor to further improve warping techniques.

Finally, we included a simple laser-pointer condition in our experiment, in which reference was unambiguous. This mode was vastly superior to any other mode, including the pointer perspective mode. Hence, the following suggestions can be derived from the data. Virtual interactions could greatly benefit of some means of unambiguous distal reference, such as a virtual laser pointer. In our opinion, this method is unlikely to be rivaled by any warping technique. However, if a CVE setting aims for realism, interactions can be improved with warping techniques. Even if the accuracy of pointing will still be limited, small reductions in pointer-observer biases may facilitate virtual interactions.

4.2 Integration of Speech and Pointing

The experiment showed that speech cannot fully compensate for errors in pointing-based communication. One possibility is, that observers did not wait for the pointers' descriptions in a subset of trials. However, the inspection of the data revealed that reaction times in speech trials that eventually resulted in an error were on average larger than in correct trials and did not differ substantially in speech content (Supplemental Table 3). Thus, it seems that errors were made despite verbal descriptions. Another hypothesis is that observers sometimes selected planets that accidentally corresponded to the pointers' descriptions. This might happen when pointers expect that observers only look for a planet in the vicinity of the pointed-at referent and hence use descriptions that single out the target only locally. Simultaneously, observers might assume that the referent must be close to their incorrect interpretation of the pointing gesture and select a planet that matches the pointer's description. In this case, the misunderstanding remains unnoticed despite speech. If that was the case, one would expect that the total errors in trials in which errors were made do not differ considerably in speech and pointing-only trials. An inspection of the data supported this hypothesis (Supplemental Table 4). This finding is also consistent with the notion that pointing gestures should be considered indicating relatively small areas of space containing the referent [10, 15] and not a precise location.

Finally, the comparison of pointing accuracy without speech and reaction times in condition with speech revealed that small changes in pointing accuracy can considerably affect the time necessary to agree on a referent. In our dataset, an improvement of pointing accuracy of 1° resulted in a 1.4 s shorter reaction time. This effect was corroborated when comparing only the naïve and warping mode. As both modes were indistinguishable for the participants, the effect cannot be attributed to situation-induced changes in the interlocutors' behaviors, such as changed level of caution or verbalization when pointing appears more or less precise [12]. This suggests that interlocutors may try to refer to specific locations or small areas with pointing gestures and not only broader regions of space. Of course, the potential gain associated with improved pointing accuracy can be expected to depend highly on the setting. We deliberately opted for a situation in which precise pointing is required and giving verbal descriptions is difficult. Thus, relatively high gains of more accurate pointing could have been expected. How these effects play out in other situations is an open question.

4.3 Conclusion

The study revealed a close relationship between pointing acuity and the speed and effectiveness of establishing a joint focus of attention. Speech could compensate for the limited acuity of pointing perception only partially and at the expense of reducing communication efficiency. Attempts to facilitate pointing-based communication in CVE can potentially have a considerable effect but are not straight forward to implement. Our simple, parameter-free warping mode indeed increased misunderstandings. This suggests that a warped pointer representation might need to consider more information, such as the observer viewpoint, individual factors, or aspects of the CVE. Nevertheless, if such advanced warping techniques could approximate the performance of our pointer perspective mode – which should be theoretically possible – pointing-based interaction with and without speech in CVEs could be improved considerably while still giving the impression of a natural interaction.

Acknowledgments. We thank Anne Hanfbauer and Stefanie Flepsen for their help with the data collection. This work was supported by the German Research Foundation DFG (Grants HE6710/5–1 and HE6710/6–1 to Oliver Herbort).

References

1. Bakdash, J.Z., Marusich, L.R.: Repeated measures correlation. *Front. Psychol.* **8**(456), 1–13 (2017). <https://doi.org/10.3389/fpsyg.2017.00456>
2. Bangerter, A.: Using pointing and describing to achieve joint focus of attention in dialogue. *Psychol. Sci.* **15**(6), 415–419 (2004). <https://doi.org/10.1111/j.0956-7976.2004.00669>
3. Bangerter, A., Oppenheimer, D.M.: Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture* **6**(1), 85–102 (2006). <https://doi.org/10.1075/gest.6.1.05ban>
4. Butterworth, G., Itakura, S.: How the eyes, head and hand serve definite reference. *Br. J. Dev. Psychol.* **18**(1), 25–50 (2000). <https://doi.org/10.1348/026151000165553>
5. Herbort, O., Krause, L.-M., Kunde, W.: Perspective determines the production and interpretation of pointing gestures. *Psychon. Bull. Rev.* **28**(2), 641–648 (2020). <https://doi.org/10.3758/s13423-020-01823-7>
6. Herbort, O., Kunde, W.: Spatial (mis-)interpretation of pointing gestures to distal referents. *J. Exp. Psychol.: Hum. Percept. Perform.* **42**(1), 78–89 (2016). <https://doi.org/10.1037/xhp0000126>
7. Herbort, O., Kunde, W.: How to point and to interpret pointing gestures? Instructions can reduce pointer-observer misunderstandings. *Psychol. Res.* **82**(2), 395–406 (2018). <https://doi.org/10.1007/s00426-016-0824-8>
8. Krause, L.M., Herbort, O.: The observer’s perspective determines which cues are used when interpreting pointing gestures. *J. Exp. Psychol.: Hum. Percept. Perform.* **47**(9), 1209–1225 (2021). <https://doi.org/10.1037/xhp0000937>
9. Louwse, M.M., Bangerter, A.: Focusing attention with deictic gestures and linguistic expressions. In: *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (2005)
10. Lücking, A., Pfeiffer, T., Rieser, H.: Pointing and reference reconsidered. *J. Pragmat.* **77**, 56–79 (2015). <https://doi.org/10.1016/j.pragma.2014.12.013>

11. Mayer, S., et al.: Improving humans' ability to interpret deictic gestures in virtual reality. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14. Association for Computing Machinery, New York (2020). <https://doi.org/10.1145/3313831.3376340>
12. Pechmann, T., Deutsch, W.: The development of verbal and nonverbal devices for reference. *J. Exp. Child Psychol.* **34**(2), 330–341 (1982). [https://doi.org/10.1016/0022-0965\(82\)90050-9](https://doi.org/10.1016/0022-0965(82)90050-9)
13. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022)
14. Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M.S.: afex: Analysis of factorial experiments (2022)
15. van der Sluis, I., Krahmer, E.: Generating multimodal references. *Discourse Process.* **44**(3), 145–174 (2007). <https://doi.org/10.1080/01638530701600755>
16. Sousa, M., dos Anjos, R.K., Mendes, D., Billinghamurst, M., Jorge, J.: Warping DEIXIS: distorting gestures to enhance collaboration. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12. CHI 2019, Association for Computing Machinery, New York (2019). <https://doi.org/10.1145/3290605.3300838>
17. Wnuczko, M., Kennedy, J.M.: Pivots for pointing: visually-monitored pointing has higher arm elevations than pointing blindfolded. *J. Exp. Psychol.: Hum. Percept. Perform.* **37**(5), 1485–1491 (2011). <https://doi.org/10.1037/a0024232>
18. Wong, N., Gutwin, C.: Where are you pointing? The accuracy of deictic pointing in CVEs. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2010), pp. 1029–1038 (2010). <https://doi.org/10.1145/1753326.1753480>
19. Wong, N., Gutwin, C.: Support for deictic pointing in CVEs: still fragmented after all these years. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1377–1387. ACM (2014). <https://doi.org/10.1145/2531602.2531691>