The Impacts of Referent Display on Gesture and Speech Elicitation

Adam S. Williams *Member, IEEE*, and Francisco R. Ortega *Member, IEEE*

Abstract—Elicitation studies have become a popular method of participatory design. While traditionally used to examine unimodal gesture interactions, elicitation has started being used with other novel interaction modalities. Unfortunately, there has been no work that examines the impact of referent display on elicited interaction proposals. To address that concern this work provides a detailed comparison between two elicitation studies that were similar in design apart from the way that participants were prompted for interaction proposals (i.e., the referents). Based on this comparison the impact of referent display on speech and gesture interaction proposals are each discussed. The interaction proposals between these elicitation studies were not identical. Gesture proposals were the least impacted by referent display, showing high proposal similarity between the two works. Speech proposals were highly biased by text referents with proposals directly mirroring text-based referents an average of 69.36% of the time. In short, the way that referents are presented during elicitation studies can impact the resulting interaction proposals; however, the level of impact found is dependent on the modality of input elicited.

Index Terms—Human computer interaction (HCI), User studies, Mixed / augmented reality, Gestural input, Elicitation

1 Introduction

Designing effective systems requires an in-depth knowledge of the user, their interactions, and how they think [17]. One path towards gaining that understanding is to run an elicitation study. Elicitation studies use a mix of observational and participatory design methodologies to gain an understanding of how users interact with a system [56]. Elicitation can be performed under a wide range of goals, sometimes using an emulated version of an emerging technology [43,52], conceptual technologies [7], or existing technologies [22].

This study design was integrated into human-computer interaction research by Wobbrock et al. in 2005 [62] and later popularized by the same team [63]. Self-described as a "guessability study", Wobbrock et al.'s goal was to find inputs that were discoverable to new users of a multi-touch system [62]. By observing users interact with a system in which the gulf of execution (e.g., barriers of execution) has been removed, that user's natural behaviors and interactions can be observed. While these interactions will vary from user to user, an aggregation of multiple users' interactions can be used to derive a consensus set of discoverable (i.e., guessable) interaction proposals.

While generating a consensus set is often a major goal of elicitation studies, it is not the only possible outcome. Discoveries beyond a consensus set can be made by interpreting the observational data collected during the study. Examples of this type of finding include the impact of scale on interaction generation [41, 50], the timing information around co-occurring gesture and speech inputs [28, 60], user modality preference when multiple modality options are available [11, 34, 36], and that users have an prefer multimodal interactions more as task cognitive load increases [40].

The popularity of elicitation studies is evident in the number of domains that have performed them. These domains include multi-touch surfaces [5, 32], mobile devices [46], mid-air gestures [38, 43, 59], television browsing [34, 36], computer-aided design [22], and internet of things home sets ups [65]. Researchers have further adapted elicitation methodologies to examine a range of interaction paradigms from using multi-touch and mid-air devices in tandem [44, 61] to imposing constraints on the users' motion to investigate gesture sets suitable for both impaired and non-impaired users [1,47]. As of the year 2020 More than 216 elicitation studies have been run, totaling to 5,458 participants, and 3,625 commands (i.e., referents) tested [56].

Alongside the widespread use of elicitation methodology comes a stream of modifications and improvements upon it. Ten years after the original paper, the "Agreement Index", a metric of proposal consensus, was improved and became the "Agreement Rate" [54, 62]. Other changes to the calculation of consensus include between groups metrics [54], production agreement [53], dissimilarity of proposals metrics [53], the addition of speech proposal consensus metrics [34], and statistics to help verify the prevalence of chance agreement [51]. Some studies directly emulate the work of Wobbrock et al. [56, 63], while others radically alter the process [6]. There have been variations of the Wizard of Oz systems used [9], the presentation of referents [56], and even attempts to deliberately prime users with a certain mindset [7, 47] or mental frames [6].

Elicitation has most often been used to derive interactions that use gesture as the input modality [56]. These gestures may be limited to a single body part (i.e., hands) [63] or use combinations of body parts [53]. As new technologies continue to emerge, elicitation is starting to use input modalities outside of gesture. Examples of this shift are most commonly seen in studies examining gesture and speech based inputs [22, 34, 36, 59, 60].

An area that has been unquestioned in the literature today is, "Does changing the way referents are presented impact elicited interaction proposals" Herein lies a concerning facet in this ever-evolving body of literature; there is a scarcity of work examining how minor changes to elicitation design can impact the resulting interactions elicited. To begin answering that question this paper presents a comparison of two gesture and speech elicitation studies done for basic object manipulations in optical see-through augmented reality (AR) environments [59, 60]. The difference in elicitation design between these studies is limited to the display of referents (e.g., commands being elicited).

Out of that comparison this work uncovers evidence of a biasing effect in elicitation caused by participants imitating referents as they were displayed. This work discusses how referent display impacted the elicited proposals for both gesture and speech input modalities. Finding that speech was more susceptible to referent biasing than gestures. Lastly, design recommendations are provided to help mitigate the impacts of referent biasing.

2 Background

Most elicitation studies follow a similar protocol. Commonly around 25 (Median = 20 Standard Deviation (SD) = 4) participants are recruited [56]. These participants are then asked to generate input proposals for a list of referents to be executed. These referents are presented one at a time while the participant produces an input proposal they feel is appropriate for that referent using the input modality requested.

- *Adam S. Williams is with the Colorado State University Computer Science Department. E-mail: AdamWil@colostate.edu.*
- *Francisco R. Ortega is with the Colorado State University Computer Science Department. E-mail: F.Ortega@colostate.edu.*

Manuscript received 11 March 2022; revised 11 June 2022; accepted 2 July 2022. Date of publication 01 September 2022; date of current version 03 October 2022. Digital Object Identifier no. 10.1109/TVCG.2022.3203090

Elicitation studies often use Wizard of Oz (WoZ) experiment design which is a way to remove the gulf of execution between the participant and the system by having the experimenter trigger the recognition of inputs [63]. The referents that are used by the study are commonly specific to one domain or application.

Data is often collected using video recordings [56, 57]; however, other formats have been used [36, 53]. Videos are then hand-annotated by one or more raters and broken into gesture proposals [39, 59, 60, 63]. These will be very granular gestures with notes on features including the number of fingers used, hand position, and direction of movement [43, 60]. The granular gestures are binned into equivalence classes based on predefined similarity features or insights from previous work. Binning proposals is an important step towards removing the individual-level characteristics of the proposals in favor of a more generalizable consensus set. Lastly, the annotated data is paired with the observational data from interviews and participant commentary.

Agreement metrics are used to quantify consensus across participants and referents using the binned gesture proposals. The most popular metric for agreement is *Agreement Rate*, which is a measure of pairs of participants in agreement over all possible pairs [63]. Other metrics such as machine learning techniques [53] and metrics designed for speech inputs [34] exist. Based on the metrics are used, a set of consensus gestures is proposed for referents that achieve higher than a predetermined level of agreement. Design guidelines informed by the study's proposal space and observational data can serve as additional contributions of elicitation studies.

2.1 Imitation

This paper raises the issue of imitation as a concern to be addressed when designing an elicitation study. Imitation is a natural human trait that is deeply ingrained in everyday social and physical processes. Imitation of visible action has been considered either an inborn skill [31] or learned via self-observation [4] and reinforced from a young age [13]. Regardless of where it arises from, the existence of action imitation is the same. Non-human representations (i.e., a wooden hand) stimulate lower imitation than human representations [29] but geometric objects seem unaffected by this if their action can reasonably be completed by a human [16], as is the case for most referents (i.e., move left). Separately, speech imitation is a skill used from a young age to facilitate language learning [25]. When prompting participants with text/read-aloud referents, their responses may be biased towards imitating the referent as stated. Alternatively, when using images or animations of objects as referents, the properties of those representations may bias participants to imitate the shape or motion of that object.

2.2 Referent Display

The goal of presenting a referent is to establish the command to be completed by the participants interaction proposal. If eliciting commands for television-based web browsing then a referent could be *refresh page* [34]. The referent *Refresh page* could be presented as text reading "refresh page", an animation of a web page being refreshed, or as an experimenter reading the referent aloud. In the case of Morris, 2012, and Nebeling et al. 2014, it was both showing the effect of the referent (e.g., the animation) and stating its name aloud [34, 36].

Referent display techniques have included animations paired with spoken aloud instructions [34], images [8, 27, 45, 48], animations alone [14, 20, 22, 24, 30, 60], text alone [59], only read aloud [10], text and animation [15, 39], text and read aloud [38, 47, 66], and the combination of text, reading aloud, and animations [52]. This variety of presentation is concerning considering that the type of presentation used could impact the study's interaction proposals.

3 Related Work

There have been limited works that use similar enough elicitation designs to be reasonably compared [56]. This section compares two such pairs of elicitation studies to provide more information on how minor changes in methodology can impact interaction proposals. These changes were not necessarily to the referent display.

3.1 Study pair 1: pen+multi-touch interactions

The first comparison is between Sukumar et al. (2018) [49] who replicated the work of Wolf et al. (1987) [64], and Welbourn et al. (1988) [58]. This study elicited pen and touch-based gestures on a multi-touch surface for use in text editing applications [49]. Sukumar et al. used a modified elicitation methodology based on the work of Wobbrock et al. in 2005 [49, 62]. Note that while the works of Wolf et al. [64] and Welbourn et al. [58] occurred before 2005 they were observational studies that used methods similar to elicitation.

Both of these studies observed participant behavior during a writing and text editing task. The main difference between these works was the use of a multi-touch surface [49] as opposed to pen and paper [58, 64]. That difference was further pronounced by telling participants they were interacting with a live recognition system compared to paper alone.

The study employing multi-touch devices found some interactions to be quite similar to the prior two studies, examples being the gestures proposed for the referents "insert", "delete", and "move". Sukumar et al. found differences in the referents "join", "split", and "new paragraph" which were conceptually similar to the commands used in the previous experiment [58], but had a different wording [49]. They go on to speculate that the differences in results were caused by those variations in terminology, citing other work that used the same terms to produce similar results during multi-touch elicitation [12]. The differences in terminology used are akin to differences in text or spoken referents. Additionally, the differences in the participant's experiences with technology caused by 30 years of technological advancement likely contributing factors to differences in results [35].

3.2 Study pair 2: gesture+speech interactions

Nebeling et al. (2014) conducted a gesture and speech elicitation study following the design used by Morris (2012) [34]. These studies elicited gesture and speech commands for a television-based web browser equipped with a Microsoft Kinect. Each study used a sample size of 25 participants who were grouped in pairs with a single triad and asked them to generate either a speech, gesture, or gesture+speech command for each referent. The referents were shown as animations and read-aloud. The work of Nebeling et al. replicated the conditions of Morris, 2012 [34] as closely as possible, omitting only a few of the original referents.

Participant's interaction modality preferences were largely the same between the two studies, choosing to use either speech alone 56% (Morris, 2012) and 65% (Nebeling et al.) of the time, gesture alone 41% and 31% of the time, or multimodal gesture+speech interactions 3% and 4% of the time [34, 36]. More varied results are found in the interaction proposals. Each study had some overlap between proposals but differing proposal frequencies. An example of this is seen in the proposals for the referents "go back" which had 7 participants propose "flick hand (arrow)" in Morris' study and 5 in Nebeling et al.'s study. Some referents had less similarity, demonstrated by the "click link" referent which had 7 "hand-as-mouse + click/grip" proposals in Morris, 2012, and 11 in Nebeling et al. Differences in past exposure and the demographics of the participants may have contributed to the variation in results. Most participants in the original work had some exposure to the Microsoft Kinect whereas very few participants had that exposure in the later study.

These studies elicited speech interactions by using referents that were both read out loud and animated for participants. While neither paper directly addresses the impact that reading referents out loud had on their results, the impacts can be seen for some referents where the most common speech proposal was identical or nearly identical to the referent as it was read. Examples of this imitation are proposing "open browser" for the referent *open browser* [34] or "go back" for the referent *go back* [36]. Not all referents were repeated and participants could propose other modalities of input, lessening the impacts of any biases introduced by reading the referents. For example, while saying "open browser" was the second most common proposal for the referent *open browser*, the top proposal was a gesture where people acted out pressing the open browser button [34].

4 METHODS

To further explore the impacts of elicitation design changes, this work compares two recent studies using the annotated data from those works as provided by the authors [59, 60]. These studies observed participant's interactions and behaviors while completing basic tasks within a generic AR environment. The input modalities examined in these studies were mid-air gesture, speech, and the combination of mid-air gesture+speech [59, 60]. These two experiments were similar apart from the way the referent was displayed. Both experiments were run on a Magic Leap One optical see-through augmented reality head-mounted display using the same software, apart from the referent display changes. Both works used video to capture the raw participant interactions and were performed by the same researchers, in the same room, using the same subject pool, with no subject taking part in both studies. The first study used text referents (top of Figure 1), referred to as "E-Text" [59]. The second study used animated referents with no text shown except the modality to be used (bottom of Figure 1), referred to as "E-Animated" [60].

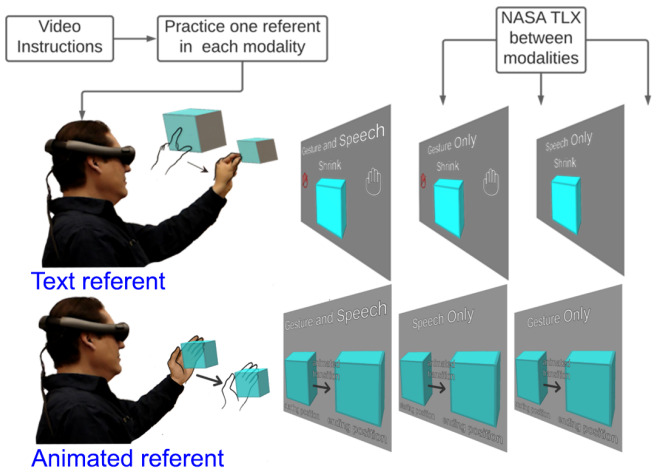

Fig. 1. High level study flow, Top: text referent (E-Text) [59], Bottom: animated referent (E-Animated) [60]

4.1 Compared Work's Designs

The two experiments examined here each used a WoZ design. Participants were videotaped while interacting with the system. The participants' inputs were only constrained by the input modality of the condition that they were in. Within each input modality condition, participants were invited to generate any input proposal that they felt was appropriate for the referent presented. For example, if the modality was speech then any utterance proposed was accepted causing the experimenter to trigger the system's response to that input, thus advancing the experiment.

In both experiments, participants first completed the informed consent and demographics questionnaires. The demographics questionnaire was used to establish the participant's previous exposure to mid-air gestures (e.g. the Microsoft Kinect), virtual reality (VR), and AR. This questionnaire also included standard demographic questions such as age, gender, and handedness attributes.

Next, the participants viewed an instruction video that explained the experiment. These video instructions were similar for both studies apart from referent presentations (i.e., animated with E-Animated, text with E-Text). The videos outlined the high-level objectives of the experimental tasks. Participants were given a practice round where they generated one input proposal per input modality for a color change referent.

After the practice round, participants were shown interaction modalities in a counterbalanced order. Within each modality, the referents were shown in random order. In each trial the participant was shown a cube that was centered in their viewport and rendered approximately 50 cm away. The NASA Task Load Index (NASA-TLX) survey was administered after the completion of all referents for a given input modality condition to measure that input modalities perceived workload [18].

4.1.1 Differences in Methods

In E-Text, the referents were shown as text and read aloud (top of Figure 1) [59]. Participants were told they were interacting with a live system, leveraging the Wizard-of-Oz design. Upon initiation of a proposal, the experimenter would trigger the system's recognition of that input which would then execute that referent's animation. The animations ran for two seconds, then a blue screen was shown. After another delay, the next referent was loaded. This cycle would continue until all referents and modalities were completed.

In E-Animated, the referents were shown as animations that were triggered two seconds after loading the cube [60]. Participants would see a blue screen, then the rendered cube and modality information (right of Figure 1). After a two-second delay, the referent would execute the same animations as shown in E-Text with exceptions to the abstract referents which used different animations. In E-Text, these animations were shown after a proposal was made, and in E-Animated, they were shown before. Upon seeing the animations, participants had to "guess" what command a fictitious participant in another room used to generate that input proposal fostering a belief that the system was live but disabled for them. That design choice was made to try to capture feelings of interaction with a live system as seen in E-Text [60].

With either design, the authors of those works expected differences based on the level of priming caused by either the animations or the text used [59, 60]. When the referents were shown as text, the proposed speech was expected to closely follow that text. The use of text referents in elicitation is common [10, 38, 47, 52, 66]. When prompting the user with animations, the gestures produced might be primed by the movements of the objects. This design is also common within elicitation studies [14, 20, 22, 24, 30, 60]. In the few elicitation studies that allowed speech inputs the impact of referent display on proposals was never stated [34, 36].

4.1.2 Referents Used

These studies used referents that are considered canonical manipulations for generic interactions with 3-dimensional user interfaces [3, 37]. The canonical referents used were scaling, translation on each axis, and rotation about each axis. In addition to those referents, the abstract referents of selection, create, and destroy were used [59, 60].

4.1.3 Participants

Each study consisted of a unique set of 24 volunteers (E-Text: 4 female, 20 male; E-Animated: 10 Female, 14 Male) summing to 48 participants in total. Participants were recruited using emails and through word of mouth. Ages ranged from 18-43 years (Mean = 23.32, SD = 5.23) in E-Text and 18-46 years old (Mean = 25, SD = 6.9) in E-Animated. Two participants in E-Text and five in E-Animated were left-handed. In E-Text, eleven participants reported less than 30 minutes of stereoscopic AR usage before the experiment. In E-Animated five participants reported weekly use of VR. Only two of those participants used VR for more than 5 hours weekly (5 hours, 10 hours). The other three participants reported 1-3 hours of weekly VR use. Several participants did not learn English as a first language but reported fluency in it(E-Text: 8, E-Animated: 7). Across both experiments, all participants reported normal or corrected to normal vision.

4.2 Data Preparation

The researchers hand-annotated the collected videos to produce the data that was interpreted during those studies [59, 60]. Participants made gesture proposals for each referent in both the gesture alone and gesture+speech conditions. Participants proposed utterances for each referent in the speech and gesture+speech conditions.

4.2.1 Gesture Data Preparation

Gestures were annotated from the video at a granular level then binned into high-level equivalence classes. The same coders were used across

both studies. At the granular level gestures were binned based on fingers used, hands used, the shape of the hand, and motion of the gesture. These classes were then collapsed based on groupings of fingers used and hand poses. Some examples of these are "grasping" where all fingers were closed, "pinching" where just the thumb and index or thumb, index, and middle fingers were touching, "open" where all fingers were extended, and "index finger" where only the index finger was extended. Additionally, movements along the same axis were considered the same. For example, translations right and left were both considered movements along the y-axis. These equivalence classes were considered reasonable given that users care less about the count of fingers used than the hand pose used [61].

4.2.2 Speech Data Preparation

The utterances proposed by each participant were hand transcribed from the video recordings of the speech only and the gesture+speech conditions. The speech data was then binned based on the syntax used. These bins included words that indicated action, direction, and object specification. Some articles of speech were discarded for this analysis such that saying "move the object left" which was considered the same as "move object left". Separately the utterances proposed were grouped by common words. These groups used strict criteria where "move backwards" and "move backward" would be considered the same but "move back" would be different.

4.2.3 Gesture+Speech Data Preparation

The gesture proposals and speech proposals from the gesture+speech input condition were annotated individually from the videos following the same practices described for the unimodal gesture and unimodal speech conditions.

4.3 Analysis Performed

Prior to comparing these works some understanding of the analysis conducted by them is required. This section discuses those analyses and explains what those analysis can be interpreted as meaning.

4.3.1 Gesture Metrics

For gesture analysis, the main metric used was Agreement Rate ($\mathscr{AR}$). The formula for $\mathscr{AR}$ is shown in Equation 1. $\mathscr{AR}$ is a measure of how much participant agreement there is for each referent. In Equation 1, P is the set of all proposals for referent r, and P_i are the subsets of equivalent proposals from P [54]. Within each referent and input condition a participant could only have a single proposal. Bootstrapped confidence intervals for $\mathscr{AR}$ were constructed following the methods detailed by Tsandilas 2018 [51] and cross-checked using the AGATe 2.0 tool (AGreement Analysis Toolkit)[1].

In the compared works an $\mathscr{AR}$ of .3 was labeled as high agreement, meaning if the referent *select* achieved an $\mathscr{AR}$ of .5 then the most frequent proposal for *select* was considered to be discoverable by novice users of a system [59, 60]. Authors should determine what a reasonable level of $\mathscr{AR}$ is given their study design, prior work, and sample size [51, 54, 55]

$$\mathscr{AR}(r) = \frac{\sum_{P_i \subseteq P} \binom{|P_i|}{2}}{\binom{|P|}{2}} \quad (1)$$

4.3.2 Speech Analysis

Speech was analyzed using two metrics of agreement. The first was max-consensus ($\mathscr{MC}$). $\mathscr{MC}$ is the percent of participants proposing the most common utterance proposal [34]. If 12 participants proposed the utterance "move left" for the referent *move left*, 5 propose "left", 2 propose "move", and 1 participant proposes "sideways" then the $\mathscr{MC}$ equals 60%. The second speech metric used was the consensus-distinct ratio ($\mathscr{CDR}$). $\mathscr{CDR}$ is the percent of proposals for a referent that have over a baseline of 1 participants proposing them [34]. In the above-mentioned scenario, the $\mathscr{CDR}$ is 75%. $\mathscr{MC}$ and $\mathscr{CDR}$ were averaged across referents to gauge the general level of difference in metrics between the two studies. Only $\mathscr{CDR}$ was reported for E-Text, the $\mathscr{MC}$ values given here are derived from the data provided for that work [59].

These metrics capture the peak and spread of the proposal space [34]. If a referent has a proposal with a high $\mathscr{MC}$, that proposal is considered discoverable to novice users of this system. Alternatively, a high $\mathscr{CDR}$ means that a referent has a high amount of disagreement on the best choice of proposals for that referent. These works also also analyzed speech using each of the binned syntax's rate of use as a percentage of all syntax use [59, 60].

4.3.3 Other analysis

Outside of those analysis, this work will compare the raw NASA-TLX scores collected during those works.

5 Comparisons of Results

The following comparisons will differ from the original work's result reporting in order to more effectively show the similarities and differences found between the studies. In the original works the results were broken into gesture alone, speech alone, and gesture+speech conditions. That said, the proposed utterances and gestures from the gesture+speech condition were similar enough to the proposals for the gesture alone and speech alone conditions to merit only including the unimodal comparisons in this work. Comparisons for the gesture+speech condition are included in the appendix for completeness, but they do not show information that would be unexpected given the unimodal comparisons.

5.1 Gesture Comparisons

The following sections compare the gestures proposals from the gesture alone condition of both experiments.

5.1.1 Agreement Rate Comparisons

Table 1. Agreement rates per referent compared across the gesture alone condition of E-Text and E-Animated with absolute differences shown

	E-Text		**E-Animated**		Absolute
Referent	$\mathscr{AR}$	95% *CI*	$\mathscr{AR}$	95% *CI*	difference
Select	.84	(.63, 1.0)	.09	(.07, .21)	.75
Create	.08	(.08, .20)	.21	(.12, .44)	.13
Delete	.08	(.07, .20)	.11	(.08, .26)	.03
Move away	.55	(.42, .77)	.37	(.28, .54)	.18
Move towards	.39	(.23, .69)	.28	(.19, .46)	.11
Move left	.47	(.31, .71)	.49	(.35, .71)	.02
Move right	.43	(.35, .63)	.43	(.30, .66)	.00
Move up	.34	(.25, .57)	.49	(.36, .71)	.15
Move down	.41	(.26, .64)	.38	(.26, .63)	.03
Pitch Up	.12	(.09, .27)	.28	(.18, .51)	.15
Pitch Down	.16	(.11, .32)	.16	(.10, .34)	.00
Yaw left	.30	(.17, .56)	.25	(.14, .50)	.05
Yaw right	.22	(.13, .45)	.22	(.13, .44)	.00
Roll C	.51	(.33, .76)	.56	(.36, .84)	.04
Roll CC	.58	(.39, .84)	.39	(.23, .64)	.18
Enlarge	.28	(.22, .43)	.28	(.21, .43)	.01
Shrink	.22	(.17, .37)	.14	(.12, .25)	.08

Legend: C: clockwise, CC: counterclockwise, *CI*: Bootstrapped 95% confidence intervals, differences are shown as absolute values

While $\mathscr{AR}$ typically should not be compared across studies, the similarities in design between the two experiments and their levels of chance agreement (E-Text: .058, E-Animated: .054) make it reasonable to compare $\mathscr{AR}$ here. These design similarities include using the same video coders, subject pool, sample size, and recruitment methods. Chance agreement was calculated by taking the P_e term from Fleiss' Kappa equation for each study. All comparisons for the gesture only condition can be seen in Table 1 where $\mathscr{AR}$ is reported alongside bootstrapped 95% confidence intervals. A similar table is shown for

[1] Available at http://depts.washington.edu/acelab/proj/dollar/agate.html

the gesture+speech condition in Appendix A, Table 5. In the compared works an $\mathscr{AR}$ of .3 and above was considered high agreement [59, 60]. A $\mathscr{AR}$ of .3 may not always indicate high agreement, for more detailed information on interpreting agreement rates please see [51, 55]. Using that number as a benchmark, only the select referent had a difference of more than .3 in $\mathscr{AR}$. For 10 out of 17 referents the difference in $\mathscr{AR}$ was below .1, which we consider to be a minimal difference. Roll counterclockwise, select, and move towards were the only referents that had one study finding them to have high agreement, while the other did not.

Stark differences in $\mathscr{AR}$ were observed with the *select* referent which required an abstract animation (Table 1). The difference in $\mathscr{AR}$ for the *select* referent between experiments was 0.75. In E-Text select was largely accomplished by participants pointing at the virtual object. In E-Animated, proposals became far less consistent, likely due to varied interpretations of the referent's animation. The animations for the create and delete referent were particle effects animations that either materialized the object or caused it to disappear. Select was animated as a gradual hue change where the object would become brighter [60]. In that work, the select animation was chosen after piloting a few animations (i.e., and arrow pointing, bouncing the object). The hue change was the most correctly interpreted referent out of those piloted [60].

5.1.2 Granular Proposal Comparison

Heat-maps of the elicited gesture proposals from these two experiments were generated to provide a visual comparison of the differences in proposals across the two referent types; text versus animation. The heat-maps for the gesture alone condition are shown in figures 2 and 3. The heat-maps for the gesture+speech condition are provided in Appendix A, Figures 4, 5 and, 6 . These heat-maps do not list any gestures that were proposed a single time in order to reduce visual clutter. With those cases removed, column totals do not all sum up to 24.

In these heat-maps the y-axis provides a short description of the gesture proposals and the x-axis lists the referents across each of the two experiments. The individual cells represent the frequency of proposals for a given gesture with darker cells representing increased proposal frequency. In addition to that coloring, the count of proposals is provided in each cell. As an example, the first two columns of Figure 2 show the gesture proposals and their frequency for the referent *move up* in E-Animated and E-Text respectively. The high level of similarity between those columns suggests that there is little difference in the binned gesture proposals elicited with text referents compared to those elicited using animated referents.

The gesture proposals for the translation referents and the rotation referents (Figure 2) were often quite similar. The difference for most referents was a slightly increased variety of gestures proposed, as exhibited by the minor increase in the number of distinct proposals in that referent's column. The referent *pitch down* in Figure 2 provides an example of this where E-Text elicited 3 distinct proposals and E-Animated elicited 5.

The referent *select* has the largest deviation in gesture proposals between experiments (middle pair of columns in Figure 3). In E-Text there was one elicited proposal that occurred 22 times where in E-Animated the most frequent proposal slot was tied with 5 occurrences each. This discrepancy is likely caused by participant misinterpretation of the animation for the *select* referent [60]. The referent *delete* elicited a few different proposals when comparing across experiments; however, these were minimal with the largest deviation being that 7 participants proposed the "bloom" gesture in E-Animated where none proposed it in E-Text. The last referent displaying notable differences was *create* which elicited 11 proposals for "bloom" in E-Animated and only 3 in E-Text.

These heat-maps are evidence that the differences in gesture proposals for most referents are relatively minor. Abstract referents were more impacted by the shift from text to animated referents; however, the magnitude of the difference varied dependent on the relation of the animation used to the meaning of the word used in E-Text. The limited differences in proposals for *delete* indicate that it's animation was similar to the concept of the word delete where the differences in *select* show the opposite (Figure 3).

5.2 Speech Comparisons

The following two sections compare the speech proposals between the two elicitation studies. The first comparison is between the rates of syntax use. The second comparison is of the actual utterances proposed by participants and their associated agreement metrics.

5.2.1 Syntax Usage Comparisons

Syntax use was similar between the experiments for the speech condition with no difference being larger than 4% (top half of Table 3). The largest observed difference in syntax was a shift of nearly 10% between using only an action phrase and using only a direction phrase during the gesture+speech condition. Due to this shift, Table 3 shows the syntax rates for the speech alone and the gesture+speech condition of each study. In E-Text <direction> phrases alone were used 11.76% of the time and <action> phrases alone were used 28.19% of the time. In E-Animated there was nearly a 10% shift in syntax use, where <action> phrases alone were used 38.48% of the time and <direction> phrases alone were used 1.72% of the time. This is evidence that animated referent display may be more likely to elicit an action phrases where text is more likely to elicit a direction phrase with the caveat that this difference was only found in the gesture+speech condition where <direction> phrasing was more common. Looking at E-Text only, <direction> phrases alone were used 6.13% of the time during the speech condition and 11.76% of the time during the gesture+speech condition. Across other syntax categories referent display showed minimal impact with most differences being less than 4%.

5.2.2 Speech Proposals and Agreement Metrics Comparisons

The most common speech proposals in E-Text were always the referent as it was displayed (Figure 2). The information for the gesture+speech condition can be found in Appendix B, Table 6. The $\mathscr{MC}$ for those proposals was uncommonly high in every case, likely caused by participants imitating or repeating the referent. The results from E-Animated show more variety in top proposals. For the translations, the top proposal was still the referent as it would have been displayed in E-text indicating that text biasing may matter less for simple referents.

The largest difference in $\mathscr{MC}$ between the two studies was 66.67% in the speech only condition. The average difference in $\mathscr{MC}$ between studies was 42.45%. The smallest difference in $\mathscr{MC}$ was 16.67%. These numbers imply that while in some cases the difference in $\mathscr{MC}$ between referent displays may be lower, for most referents the differences were more severe. E-Text had an average $\mathscr{MC}$ of 75.26% where E-Animated had an average $\mathscr{MC}$ of 32.81%. Meaning that on average, speech proposals reported under E-Text were agreed upon by more than two-thirds of participants while speech proposals under E-Animated were agreed upon by less than a third of participants. The proposals that repeated referents in E-Text and the differences in $\mathscr{MC}$ between the studies are strong evidence that text primed users' speech proposals and that text referents lead to inflated $\mathscr{MC}$ values.

The $\mathscr{CDR}$ between these two studies was also varied. Often the $\mathscr{CDR}$ in E-Text was higher than in E-Animated meaning that E-Text had a narrower distribution of speech proposals compared to E-Animated. These results match the differences that would be expected when referents shown as text are imitated by participants. With most participants repeating the referent as shown, the diversity in the resulting proposal space was lessened (0.66 $\mathscr{CDR}$). Alternatively, in E-Animated where no text was shown, there was a much more varied space of speech proposals (0.42 average $\mathscr{CDR}$). These differences in $\mathscr{CDR}$ are further evidence that text based referents can impact the speech proposals generated during elicitation when compared to animated referents, resulting in a less diverse speech proposal space.

The abstract referents in E-Animated were negatively impacted by the referent display used. For *create* and *delete* the top proposals were "appear" and "disappear" which were similar to the referent in concept but closer to the animation used in actuality. Similarly, *Select* had a top proposal of "change" which was much further from

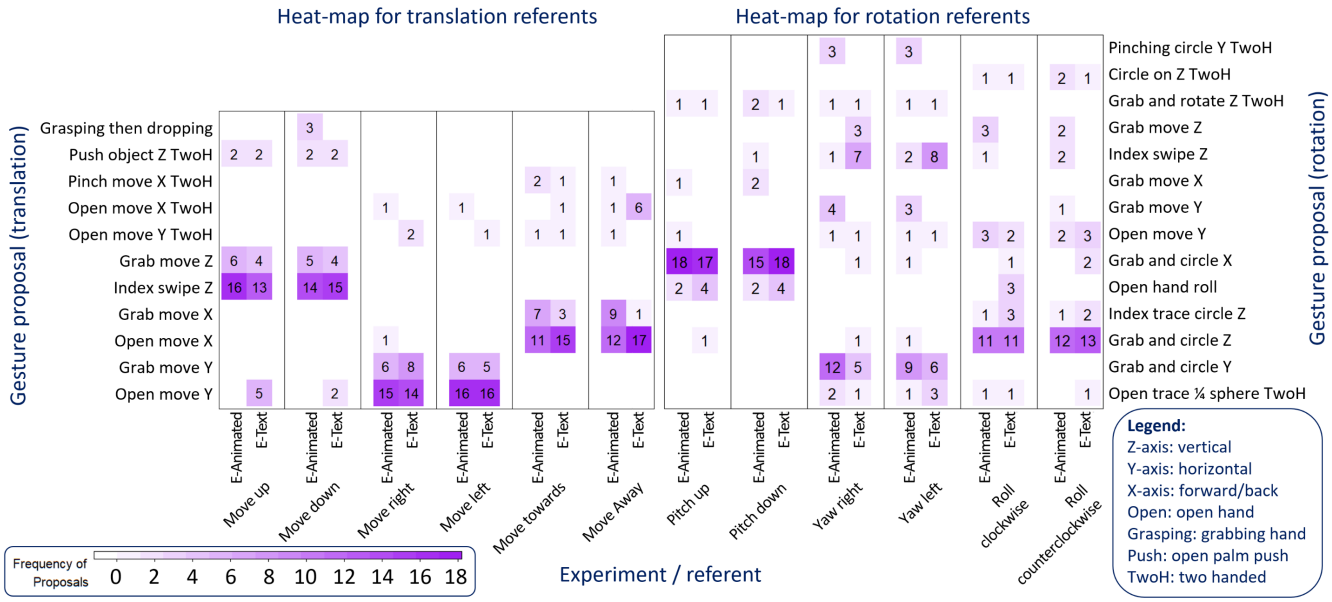

Fig. 2. Gesture proposal heat-maps by referent and experiment for translation (left) and rotation (right) referents

Table 2. Speech proposal comparisons by input condition and experiment with absolute differences and column averages

	E-Text			E-Animated			Difference	Difference
Referent	Top proposal	$\mathscr{MC}$	$\mathscr{CDR}$	Top proposal	$\mathscr{MC}$	$\mathscr{CDR}$	$\mathscr{MC}$	$\mathscr{CDR}$
Create	create	75%	0.33	appear	41.67%	0.18	33.33%	0.15
Delete	delete	91.67%	0.92	disappear	50%	0.57	41.67%	0.35
Enlarge	enlarge	66.67%	0.67	enlarge	37.5%	0.36	29.17%	0.31
Move away	move away	54.17%	0.42	move back	25%	0.38	29.17%	0.04
Move down	move down	79.17%	0.58	drop	33.33%	0.44	45.84%	0.14
Move left	move left	87.5%	0.71	move left	37.5%	0.44	50 %	0.27
Move right	move right	87.5%	0.75	move right	41.67%	0.44	45.83%	0.31
Move towards	move towards	37.5%	0.38	move forward	20.83%	0.36	16.67%	0.02
Move up	move up	79.17%	0.67	move up	54.17%	0.33	25%	0.34
Pitch down	pitch down	79.17%	0.79	rotate	20.83%	0.46	58.34%	0.33
Pitch up	pitch up	79.17%	0.75	rotate away	16.67%	0.5	62.5%	0.25
Roll C	roll C	70.83%	0.62	spin right	20.83%	0.5	62.5%	0.25
Roll CC	roll CC	70.83%	0.67	spin left	25%	0.4	50%	0.12
Select	select	87.5%	0.79	glow	20.83%	0.54	48.83%	0.27
Shrink	shrink	83.33%	0.75	shrink	45.83%	0.25	66.67%	0.25
Yaw left	yaw left	75%	0.79	spin left	33.33%	0.62	37.5%	0.5
Yaw right	yaw right	75%	0.79	spin right	29.17%	0.78	45.83%	0.01
Column average		75.26%	0.66		32.81%	0.42	42.45%	0.24

Legend: C: Clockwise, CC: Counterclockwise, $\mathscr{MC}$: Max-Consensus, $\mathscr{CDR}$: Consensus-Distinct Ratio, differences are absolute values

the referent while still close to the animation used. These results can be summarized by saying that using text referents during speech elicitation will yield proposals that are similar to the referents as they were displayed. Animated referents remove that effect but introduce interpretation issues with referents that do not have a clear animation.

5.3 NASA Task Load Index

The NASA-TLX scores for each condition of E-Text and E-Animated are shown in Table 4. The scores from the gesture+speech condition are shown here due to their differences from the gesture alone and speech alone conditions. The NASA TLX results by condition and experiment were normally distributed based on the results of Shapiro-Wilk tests: E-Animated gesture: $W(24) = .948, p = .243$, E-Animated speech: $W(24) = .967, p = 0.591$, E-Animated gesture+speech: $W(24) = .934, p = 0.117$, E-Text gesture: $W(24) = .979, p = .876$, E-Text speech: $W(24) = 0.933, p = .113$, and E-Text gesture+speech: $W(24) = 0.932, p = .105$. Welch Two Sample T-Tests support that the scores have a different mean for the gesture and gesture+speech conditions across the two studies ($t(45.537) = 2.248, p = 0.029$ and $t(43.807) = 2.893, p = 0.006$ respectively). This difference was not found in the speech condition ($t(45.761) = 0.147, p = 0.884$).

E-Animated had lower perceived workload than E-Text for each condition (Table 4). The gesture+speech condition had the largest difference in perceived workload between studies (12.6) followed by the gesture condition with a difference of 9.2. The lowest difference found was between the speech conditions with an absolute difference of 0.7. The difference in perceived difficulty in both the gesture alone and the gesture+speech condition provides evidence that participants found generating gesture proposals easier when shown animations compared to text. This difference in perceived difficulty may be caused by the ease of imitating an animation's action. Speech scores were not impacted by the choice of referent display which was unexpected as participants imitated text an average of 69.36% of the time (Max: 91.67%, Min: 37.5%) during the speech condition of E-Text.

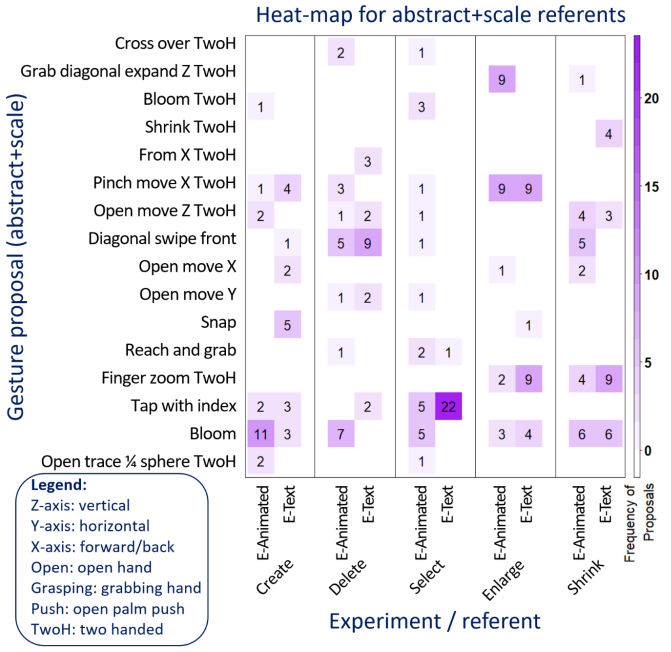

Fig. 3. Heat-map of common gesture proposals by referent and experiment (abstract and scale referents only)

These results support that there were differences between the total perceived workload between the two studies for the gesture and the gesture+speech groups. These differences were not present in the speech only conditions.

6 Discussion

These comparisons show that the choice of referent display impacted the results of these works, although, that impact was minimal when looking at gesture proposals in isolation.

Summary of differences between E-Text and E-Animated:

- **Gesture** $\mathscr{AR}$ - minor shifts in $\mathscr{AR}$ values, seen most in abstract referents
- **Gesture proposals** - minor differences in proposal frequency with abstract referents exhibiting higher variations in proposals
- **Speech syntax** - minor differences apart from a shift of approximately 10% in the use of <action> and <direction> phrasing in the gesture+speech condition
- **Speech** $\mathscr{MC}$ **and** $\mathscr{CDR}$ - large differences between studies, E-Animated had lower $\mathscr{MC}$ and $\mathscr{CDR}$ across the board
- **Speech proposals** - large differences in the utterances proposed
- **NASA-TLX** - indications that people perceived lower workload with animated referents, most notably in the gesture and gesture+speech conditions.

6.1 Referent Biasing Through Imitation

Prompting with text referents biased participants to imitate that text as part of or as the entirety of their proposal, biasing the results to be in favor of the displayed referent names. This bias artificially inflated the consensus of speech proposals. These differences were more prevalent and of a larger magnitude during the speech condition than the gesture condition. If imitation biased speech proposals are implemented into a system that does not display the same text as the referents that were used, those elicited speech commands will be far less discoverable than the elicitation study's reported $\mathscr{MC}$ would suggest.

These differences extend beyond the individual proposals. The syntax used in speech proposals also changed based on the type of referent display; however, there was an association between the syntax used across the studies. Animations caused a higher occurrence of <action> phrases compared to an increase in <direction> phrasing when using

Table 3. Frequency of syntax used across experiments by condition with absolute differences

Speech only	E-Text	E-Animated	Difference
<Action>	24.75%	28.19%	3.44%
<Action> <Direction>	50.25%	47.06%	3.19%
<Action> <Object> <Direction>	12.75%	14.22%	1.47%
<Action> <Object>	5.64%	9.31%	3.67%
<Direction>	6.13%	1.23%	4.9%
<Other>	0%	0%	0%
Gesture+speech	E-Text	E-Animated	Difference
<Action>	28.43%	38.48%	10.05%
<Action> <Direction>	43.87%	39.95%	3.92%
<Action> <Object> <Direction>	10.54%	12.99%	2.45%
<Action> <Object>	4.41%	6.86%	2.45%
<Direction>	11.76%	1.72%	10.04%
<Other>	0.98%	0%	0.98%

Legend: E-Text: text referent, E-Animated: animated referent, differences are absolute values

Table 4. NASA-TLX overall score for each experiment and condition with absolute differences shown

Gesture only condition

	E-Text	E-Animated	Difference	P-value
Mean	39.3	30.1	9.2	.029
SD	13.4	14.8	1.4	

Speech only condition

	E-Text	E-Animated	Difference	P-value
Mean	33.5	32.8	0.7	.88
SD	15.6	14.5	1.1	

Gesture+speech condition

	E-Text	E-Animated	Difference	P-value
Mean	43.5	30.9	12.6	.006
SD	13.3	16.7	3.4	

text, suggesting that observing movement may prime more consideration of the type of movement seen whereas text primes consideration around the direction that it should move. This effect was only observed in the gesture+speech condition which may be due to the way that gestures could replace utterances during that condition. In the speech condition participants could only propose utterances, potentially forcing a different use of syntax.

Gestures were often biased such that a participant would attempt to imitate the exact motion of the animation in their gesture proposal. For rotations, this looks like a gesture proposal that tries to mirror the specific degrees of rotation through the movement of the participants' wrist (Figure 2). The differences in scaling gestures were more pronounced. Users prompted with text proposed more gestures that were informed by the legacy "zoom in" and "zoom out" gestures currently used on touch-screen devices. The animation for scaling had one corner of the rendered cube fixed while the others moved outwards for a uniform expansion giving the visual effect of a diagonal movement up and towards the participant. In E-Animated, animation bias manifested as gesture proposals that had a similar formation as E-Text but used a diagonal movement where E-Text was commonly only on one axis (Figure 3).

Previous work exhibits similar indications of referent imitation. The animations used in prior work are often not directly specified so they are assumed to be a logical presentation of the referent (e.g., *move left* translates the object left over time). The scale gesture found by Khan et al. [22] match the diagonal motion found in E-Animated [60]. Imitation is also inherent in the foot gestures that presumably mirror

the movements of the animation used by the avatar in Felberbaum et al.'s work [15], or the direct manipulations for rotation and translations found by Piumsomboon et al. [43].

The effect of imitation is more directly observable in speech elicitation. In Morris, 2012, some of the referents received speech proposals that were the referent as it was read aloud [34]. "Open new tab" was proposed for *open new tab*, and "open browser" for *open browser* [34]. Nebeling et al.'s replication of Morris' work found similar imitations such as "zoom in" for *zoom in* or "go back" for *go back*. In those studies, the participant could choose between gestures, speech, or gestures+speech as input modalities making any imitation of speech less likely to inflate the $\mathscr{MC}$ scores.

6.2 Implications for Elicitation Studies

The differences in gesture proposals caused by referent display are most salient in the distance traveled by the gesture but not in the shape and general motion of the gesture. This is demonstrated by the two-handed scaling gesture encountered in these studies. In E-Text the gesture was performed along a horizontal plane in contrast to moving at an angle 45 degrees away from that plane in E-Animated. In either case, the scale gesture used open hands that extended from a central location away from each other. More support for this conclusion is seen in the rotation gestures which were often a pointing or pinching gesture that traced a circle in the air. The difference caused by animation was in the amount that a participant rotated their hand while the shape and motion of the gesture remained consistent. This information is easily lost when aggregating the gesture data from the granular bins to the equivalence classes used when computing $\mathscr{AR}$.

The top gesture proposals were the same across studies for 11 of the 17 referents. In the other cases the top proposal in one study would also have a high number of occurrences in the other. Overall, the spread of proposals overlapped heavily, evident in the heat-maps for the gesture condition (Figures 2 and 3). The bulk of the similar and higher frequency (darker) proposals occurring in both studies show this overlap. While these top proposals occur in each study, the relative frequencies of their occurrences were different. Most elicitation studies recommend aliasing the most common commands [34, 43, 59, 60, 63], as opposed to recommending the most frequent proposal for each referent (e.g., a single consensus set). Through aliasing, the overlapping proportions of the proposal space can be captured, offsetting impacts of referent biasing. These limited effects of referent display on elicited gesture proposals are beneficial as prior work that focused on gesture elicitation was likely minimally impacted by their choice of referent display.

The results of the few prior works that performed speech elicitation show some evidence that the proposals were impacted by the referent display [34, 36, 59]. Focusing on the works compared here, speech proposals were highly impacted by referent display. This was seen when the highest $\mathscr{MC}$ proposal for each referent in E-Text was the referent as it was displayed. While some of the top proposals were the same between studies, these were limited to the proposals for the translation referents. This similarity may have been due to the simplicity of the referent. One the other hand, using animated referents impacted speech proposals when the referent did not have a clear animation to use. This was seen most in the *select* referent where the most common proposal in E-Animated was "glow" instead of "select"

Speech elicitation faces two disadvantages; showing text causes an artificial increase in consensus metrics (Table 2) and showing animations can encourage proposals that deviate from the intended referent ("highlight" for *select*). As elicitation continues to be used for novel inputs outside of gesture alone the impacts of referent display need to be considered.

6.3 Alias Commands

Redundantly mapping interaction techniques to commands (aliasing) is a technique for capturing a larger group of novice user's first choice interactions [34, 59, 60, 63]. The differences in results seen here are focused on the top choice and least common proposals. Often the gesture proposal space overlapped. Aliasing, the top N gesture proposals found in elicitation studies can help to counteract the impacts of referent biased proposals. The best way to allow future designers to alias proposals is to report more than a single consensus set. Examples of this include reporting the top few proposals [34, 36], reporting proposals with hand variations included [43, 60], and showing large portions of the proposal space (Figures 2 and 3).

6.4 Referent Guidelines

As elicitation use continues to gain popularity, creating referent displays that allow for unbiased input generation is critical. To simultaneously remove the bias from multiple input modalities we recommend a goals-based elicitation method where instead of showing referents as granular commands (i.e., *select*, *move left*, *deselect*), they are displayed as high-level goals (i.e., *construct a staircase out of these objects*). This approach conveys a goal to the participant without suggesting the granular commands necessary to complete it. Using that approach the steps the user completes can be decomposed to action/interaction pairs. With that a simple movement interaction might be composed of selecting, translating, and deselecting an object, thus giving proposals for three referents. This approach removes the bias caused by explicit referents. Similar methods of observing goal completion as opposed to granular action/interaction pairs have been used in information visualization studies [2, 26]. As second approach for mitigating imitation bias we recommend adding a time delay between referent presentation and proposal generation. This approach leverages how the chance of imitation decays over time [33]. Referent-less elicitation has also been recommended as a possible elicitation approach [56]. An example of this approach is to ask users to self-report their tasks and means of achieving those tasks to inform interaction design without the use of referents [19].

The more similar the intended use case and the elicitation study are, the less the impact of imitation will matter. If eliciting commands for a system that has text icons on menus, using the same names for the referents would facilitate more transference of the proposed interactions from the elicitation study to the intended system's use.

6.5 Cultural Biasing

E-Text was conducted around the release of the *Marvel - Avengers: Endgame* film. In this film, a snapping gesture was used for the removal (i.e., deletion) of half of the human population. This gesture also occurred within E-Text for both the create and delete referents but was not seen in E-Animated. During post-study interviews, two participants in E-Text noted that they chose the snapping gesture due to the Avengers: Endgame movie. While this does not mean that other participants felt the same way, its inclusion in E-Text is an example of a gesture that may have stemmed from pop culture. Culturally influenced gestures can represent a mechanism for knowledge transfer from other domains into a new environment as can also be the case with legacy biased gestures [35]. Society's growing adaptation of speech-enabled assistants (i.e., Alexa, Google) could be a source of other culturally influenced speech based interactions. As an example, "turn on the lights" and "turn on [name of item]" are both common commands within households that use these assistants.

6.6 Report Design Choices

The results of elicitation studies have included valuable insights on human behavior [10, 23, 41, 50] and some interactions found during elicitation have been implemented with positive results [21, 42]. The key to generating findings that are usable by designers is to detail the exact methodology employed. Knowing specifics of the study design will increase a practitioner's awareness of any biases that may have been introduced as a results of using that design. Additionally, knowing how well the elicitation study matches an intended use case allows practitioners to account for variances in conditions between the study and the use case.

Differences in results can emerge from a slight modification to the referent display, gaps in time between studies [49], and participant exposure to new technologies [36]. Along with the commonly identified

design choices found in publications (i.e., referents, sample size, apparatus) authors should describe the way that referents were displayed and how long they were shown.

7 Limitations and Future Work

The two studies examined in depth here each used a simple set of referents and environment [59, 60]. Using more complicated referents (i.e., "extrude object face") or using objects with varied representations (i.e., a car) may accentuate the differences found between text elicited proposals and animation elicited proposals. Future work should directly test more of the ways that imitation manifests, comparing across other types of referent display and input modalities. Some examples of these alternative referent designs are Referent-less design [19,56], goal-based referents [2, 26], and time delays between referent presentation and proposal generation. The merits of the goals-based elicitation method and the delayed elicitation method have not been established, more work is needed to see if they lessen the impacts of referent biasing.

Some of this work's conclusions were made without the use of inferential statistics and as such should be considered with caution. The authors of this work believe that the trends seen in the differences between the elicited speech proposals signal that referent display can bias elicited speech proposals. That conclusion could be more directly tested in a controlled between-subjects study, mitigating some of the statistical noise and unknown confounds that arise from cross-experiment comparisons (i.e., sample populations, temporal differences).

Another interesting line of inquiry is inspired by the results of the NASA TLX surveys where animated referents were seen as easier to generate gesture proposals for than text referents. This information could be further examined to determine if there are ways to develop adaptive interfaces that prompt users with specific formats of information to prime what input modality is used. As an example consider a user with low manual-dexterity and another user that is hearing impaired. A system might be able to prompt the user with low manual-dexterity using text to encourage speech interactions. Conversely, the user with limited hearing may prefer to interact with gestures which could be encouraged through the use of animated interaction prompts. These uses of information display were not examined here and would need to be further investigated by future work.

8 Conclusion

Elicitation design is vulnerable to biasing through action and text imitation during proposal generation. Most elicitation studies have used referent displays that may have encouraged imitation of them (i.e., animations, text). While this biasing has likely had minimal impacts on the existing body of gesture elicitation literature, caution must be had as elicitation studies move beyond gesture inputs. Most of the differences observed here were found between the elicited speech proposals where using text referents caused inflated values for $\mathscr{MC}$, $\mathscr{CDR}$, and proposals that were imitations of the text used [59, 60]. There has not been any work detailing how referent display can impact elicitation results. We hope that this paper brings awareness to that issue and encourages deeper design consideration in future elicitation studies. When less traditional inputs are elicited (speech, multimodal combinations) variations in minor aspects of the elicitation design can lead to far divergent results.

We propose using time delays between referent presentation and proposal generation or using a goals-based elicitation strategy to help mitigate referent biasing. These changes to elicitation methodologies contribute to the continual improvement of elicitation studies. Separately, detailed methodology reporting will help designers to know under which circumstances the proposed interactions will fit and where they may generalize. This context is necessary for presenting use-able elicitation results and encouraging reproducible work [56].

Acknowledgments

This work was supported by the National Science Foundation (NSF) awards IIS-1948254, IIS-2037417, CNS-2016714, CNS-2106590, and BCS-1928502. This work was also supported by the Defense Advanced Research Projects Agency (DARPA) ARO contract W911NF-15-1-0459.

References

[1] B. Altakrouri, D. Burmeister, D. Boldt, and A. Schrader. Insights on the impact of physical impairments in full-body motion gesture elicitation studies. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 5:1–5:10. ACM, New York, NY, USA, 2016. doi: 10.1145/2971485.2971502

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117. IEEE, 2005.

[3] D. A. Bowman, E. Kruijff, J. J. LaViola, and I. Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., USA, 2004.

[4] M. Brass and C. Heyes. Imitation: is cognitive neuroscience solving the correspondence problem? *Trends in cognitive sciences*, 9(10):489–495, 2005.

[5] S. Buchanan, B. Floyd, W. Holderness, and J. J. LaViola. Towards user-defined multi-touch gestures for 3d objects. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '13, p. 231–240. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2512349.2512825

[6] F. Cafaro, L. Lyons, and A. N. Antle. Framed guessability: Improving the discoverability of gestures and body movements for full-body interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174167

[7] E. Chan, T. Seyed, W. Stuerzlinger, X.-D. Yang, and F. Maurer. User elicitation on single-hand microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 3403–3414. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858589

[8] Z. Chen, X. Ma, Z. Peng, Y. Zhou, M. Yao, Z. Ma, C. Wang, Z. Gao, and M. Shen. User-defined gestures for gestural interaction: extending from hands to other body parts. *International Journal of Human–Computer Interaction*, 34(3):238–250, 2018.

[9] P. Cohen, C. Swindells, S. Oviatt, and A. Arthur. A high-performance dual-wizard infrastructure for designing speech, pen, and multimodal interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, p. 137–140. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1452392.1452419

[10] S. Connell, P.-Y. Kuo, L. Liu, and A. M. Piper. A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface. In *Proceedings of the 12th International Conference on Interaction Design and Children*, IDC '13, p. 277–280. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2485760.2485823

[11] A. Corradini and P. R. Cohen. On the relationships among speech, gestures, and object manipulation in virtual environments: Initial evidence. In *Advances in Natural Multimodal Dialogue Systems*, pp. 97–112. Springer, 2005.

[12] G. Costagliola, M. De Rosa, and V. Fuccella. A technique for improving text editing on touchscreen devices. *Journal of Visual Languages & Computing*, 47:1–8, 2018.

[13] E. Cracco, L. Bardi, C. Desmet, O. Genschow, D. Rigoni, L. De Coster, I. Radkova, E. Deschrijver, and M. Brass. Automatic imitation: A meta-analysis. *Psychological Bulletin*, 144(5):453, 2018.

[14] N. K. Dim, C. Silpasuwanchai, S. Sarcar, and X. Ren. Designing mid-air tv gestures for blind people using user- and choice-based elicitation approaches. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, p. 204–214. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2901790.2901834

[15] Y. Felberbaum and J. Lanir. Better understanding of foot gestures: An elicitation study. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173908

[16] E. Gowen, E. Bolton, and E. Poliakoff. Believe it or not: Moving non-biological stimuli believed to have human origin can be represented as human movement. *Cognition*, 146:431–438, 2016.

[17] W. J. Hansen. User engineering principles for interactive systems. In *Proceedings of the November 16-18, 1971, Fall Joint Computer Conference*,

AFIPS '71 (Fall), p. 523–532. Association for Computing Machinery, New York, NY, USA, 1972. doi: 10.1145/1479064.1479159

[18] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139 – 183. North-Holland, USA, 1988. doi: 10.1016/S0166-4115(08)62386-9

[19] K. Hinckley, K. Yatani, M. Pahud, N. Coddington, J. Rodenhouse, A. Wilson, H. Benko, and B. Buxton. Pen + touch = new tools. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, p. 27–36. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1866029.1866036

[20] L. Hoff, E. Hornecker, and S. Bertel. Modifying gesture elicitation: Do kinaesthetic priming and increased production reduce legacy bias? In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '16, p. 86–91. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2839462.2839472

[21] Y.-J. Huang, T. Fujiwara, Y.-X. Lin, W.-C. Lin, and K.-L. Ma. A gesture system for graph visualization in virtual reality environments. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 41–45. IEEE, 2017.

[22] S. Khan and B. Tunçer. Gesture and speech elicitation for 3d cad modeling in conceptual design. *Automation in Construction*, 106:102847, 2019.

[23] A. Köpsel and N. Bubalo. Benefiting from legacy bias. *interactions*, 22(5):44–47, Aug. 2015. doi: 10.1145/2803169

[24] P. Koutsabasis and C. K. Domouzis. Mid-air browsing and selection in image collections. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, p. 21–27. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2909132.2909248

[25] P. K. Kuhl and A. N. Meltzoff. Infant vocalizations in response to speech: Vocal imitation and developmental change. *The journal of the Acoustical Society of America*, 100(4):2425–2438, 1996.

[26] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, p. 1–5. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1168149.1168168

[27] L. Lee, Y. Javed, S. Danilowicz, and M. L. Maher. Information at the wave of your hand. In *Proceedings of HCI Korea*, HCIK '15, p. 63–70. Hanbit Media, Inc., Seoul, KOR, 2014.

[28] M. Lee and M. Billinghurst. A wizard of oz study for an ar multimodal interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, p. 249–256. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1452392.1452444

[29] R. Liepelt and M. Brass. Top-down modulation of motor priming by belief about animacy. *Experimental psychology*, 2010.

[30] K. R. May, T. M. Gable, and B. N. Walker. Designing an in-vehicle air gesture set using elicitation methods. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '17, p. 74–83. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3122986.3123015

[31] A. N. Meltzoff and M. K. Moore. Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental psychology*, 25(6):954, 1989.

[32] M. Micire, M. Desai, A. Courtemanche, K. M. Tsui, and H. A. Yanco. Analysis of natural gestures for controlling robot teams on multi-touch tabletop surfaces. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ITS '09, pp. 41–48. ACM, New York, NY, USA, 2009. doi: 10.1145/1731903.1731912

[33] A. Moors and J. De Houwer. Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, 132(2):297, 2006.

[34] M. R. Morris. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, pp. 95–104. ACM, New York, NY, USA, 2012. doi: 10.1145/2396636.2396651

[35] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, M. c. Schraefel, and J. O. Wobbrock. Reducing legacy bias in gesture elicitation studies. *Interactions*, 21(3):40–45, May 2014.

[36] M. Nebeling, A. Huber, D. Ott, and M. C. Norrie. Web on the wall reloaded: Implementation, replication and refinement of user-defined interaction sets. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, ITS '14, p. 15–24. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2669485.2669497

[37] F. R. Ortega, F. Abyarjoo, A. Barreto, N. Rishe, and M. Adjouadi. *Interaction design for 3D user interfaces: The world of modern input devices for research, applications, and game development*. CRC Press, 2016.

[38] F. R. Ortega, A. Galvan, K. Tarre, A. Barreto, N. Rishe, J. Bernal, R. Balcazar, and J. Thomas. Gesture elicitation for 3d travel via multi-touch and mid-air systems for procedurally generated pseudo-universe. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 144–153. IEEE, Los Angeles, CA, USA, 2017.

[39] F. R. Ortega, K. Tarre, M. Kress, A. S. Williams, A. B. Barreto, and N. D. Rishe. Selection and manipulation whole-body gesture elicitation study in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1723–1728. IEEE, Osaka, Japan, Japan, 2019.

[40] S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, p. 129–136. Association for Computing Machinery, New York, NY, USA, 2004. doi: 10.1145/1027933.1027957

[41] T. Pham, J. Vermeulen, A. Tang, and L. MacDonald Vermeulen. Scale impacts elicited gestures for manipulating holograms: Implications for AR gesture design. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 227–240. ACM, June 2018.

[42] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 73–82. IEEE, Munich, Germany, 2014.

[43] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, p. 955–960. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2468356.2468527

[44] T. Plank, H.-C. Jetter, R. Rädle, C. N. Klokmose, T. Luger, and H. Reiterer. Is two enough?: ! studying benefits, barriers, and biases of multi-tablet use for collaborative visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 4548–4560. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025537

[45] G. A. Rovelo Ruiz, D. Vanacken, K. Luyten, F. Abad, and E. Camahort. Multi-viewer gesture-based interaction for omni-directional video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 4077–4086, 2014.

[46] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction, 2011.

[47] J. Ruiz and D. Vogel. Soft-constraints to reduce legacy and performance bias to elicit whole-body gestures with low arm fatigue. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 3347–3350. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702583

[48] C. Silpasuwanchai and X. Ren. Designing concurrent full-body gestures for intense gameplay. *Int. J. Hum.-Comput. Stud.*, 80(C):1–13, Aug. 2015. doi: 10.1016/j.ijhcs.2015.02.010

[49] P. Talkad Sukumar, A. Liu, and R. Metoyer. Replicating user-defined gestures for text editing. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, ISS '18, p. 97–106. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3279778.3279793

[50] K. Tarre, A. S. Williams, L. Borges, N. D. Rishe, A. B. Barreto, and F. R. Ortega. Towards first person gamer modeling and the problem with game classification in user studies. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, pp. 125:1–125:2. ACM, New York, NY, USA, 2018. doi: 10.1145/3281505.3281590

[51] T. Tsandilas. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Trans. Comput. Hum. Interact.*, 25(3):18, June 2018.

[52] R.-D. Vatavu. There's a world outside your tv: Exploring interactions beyond the physical tv screen. EuroITV '13, p. 143–152. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2465958.2465972

[53] R.-D. Vatavu. The dissimilarity-consensus approach to agreement analysis in gesture elicitation studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300454

[54] R.-D. Vatavu and J. O. Wobbrock. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 1325–1334. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702223

[55] R.-D. Vatavu and J. O. Wobbrock. Clarifying agreement calculations and analysis for end-user elicitation studies. *ACM Trans. Comput.-Hum. Interact.*, 29(1), jan 2022. doi: 10.1145/3476101

[56] S. Villarreal-Narvaez, J. Vanderdonckt, R.-D. Vatavu, and J. A. Wobbrock. A systematic review of gesture elicitation studies: What can we learn from 216 studies. In *Proceedings of ACM Int. Conf. on Designing Interactive Systems (DIS'20)*, p. NA. ACM Press, Eindhoven, 2020.

[57] P. Vogiatzidakis and P. Koutsabasis. Gesture elicitation studies for Mid-Air interaction: A review. *Multimodal Technologies and Interaction*, 2(4):65, Sept. 2018.

[58] L. K. Welbourn and R. J. Whitrow. A gesture based text editor. In *Proceedings of the Fourth Conference of the British Computer Society on People and Computers IV*, p. 363–371. Cambridge University Press, USA, 1988.

[59] A. S. Williams, J. Garcia, and F. Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3479–3489, 2020. doi: 10.1109/TVCG.2020.3023566

[60] A. S. Williams and F. R. Ortega. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proc. ACM Hum.-Comput. Interact.*, 4(ISS), nov 2020. doi: 10.1145/3427330

[61] M. L. Wittorf and M. R. Jakobsen. Eliciting Mid-Air gestures for Wall-Display interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 3:1–3:4. ACM, New York, NY, USA, 2016.

[62] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers. Maximizing the guessability of symbolic input, 2005.

[63] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009.

[64] C. G. Wolf and P. Morrel-Samuels. The use of hand-drawn gestures for text editing. *International Journal of Man-Machine Studies*, 27(1):91–102, 1987.

[65] I.-A. Zaiţi, Ş.-G. Pentiuc, and R.-D. Vatavu. On free-hand TV control: experimental results on user-elicited gestures with leap motion. *Pers. Ubiquit. Comput.*, 19(5):821–838, Aug. 2015.

[66] I.-A. Zaiţi, Ş.-G. Pentiuc, and R.-D. Vatavu. On free-hand tv control: experimental results on user-elicited gestures with leap motion. *Personal and Ubiquitous Computing*, 19(5-6):821–838, 2015.