THESIS

THE IMPACT OF REFERENT DISPLAY ON INTERACTION PROPOSALS DURING MULTIMODAL ELICITATION STUDIES

Submitted by
Adam S. Williams
Department of Computer Science

In partial fulfillment of the requirements
For the Degree of Master of Science
Colorado State University
Fort Collins, Colorado
Spring 2021

Master's Committee:

Advisor: Francisco R. Ortega

Ross Beveridge
Julia Sharp

ABSTRACT

THE IMPACT OF REFERENT DISPLAY ON INTERACTION PROPOSALS DURING MULTIMODAL ELICITATION STUDIES

Elicitation studies have become a popular method of participatory design. While traditionally used for finding unimodal gesture-based inputs elicitation has been increasingly used for deriving multimodal interaction techniques. This is concerning as there has been no work that examines how well elicitation methods transfer from unimodal gesture use to multimodal combinations of inputs. This work details a comparison between two elicitation studies that were similar in design apart from the way participants were prompted for interaction proposals. Referents (e.g., commands to be executed) were shown as either text or animations. Interaction proposals for speech, gesture, and gesture+speech input modalities were elicited. Based on the comparison of these studies and other existing elicitation studies the concern of referent display priming uses proposed interaction techniques is brought to light. The results from these elicitation studies were not reproduced. Gesture proposals were the least impacted. With high similarity in the overall proposal space. Speech was biased to have proposals imitating the text as displayed an average of 69.36%. The time between gesture and speech initiation in multimodal use was 166.51% longer when prompted with text. The second contribution of this work is a consensus set of mid-air gesture inputs for use with generic object manipulations in augmented reality environments. This consensus set was derived from the elicitation study that used text-based referent displays which were found to be less biasing on participant gesture production than the animated referents.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Francisco R. Ortega, for his support, encouragement, and mentorship. I also would like to thank my M.S. thesis committee members, Dr. J. Ross Beveridge and Dr. Julia Sharp for offering their time and advice. This work was supported by the National Science Foundation (NSF) awards: IIS-1948254 and BCS-1928502.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Designing usable systems requires an in-depth knowledge of the user, their interactions, and how they think [2]. A recent and popular technique used to gain this understanding and to guide interaction design is to run an elicitation study [3]. Elicitation studies are a way to observe unconstrained user behavior. Often this is done with an emulated version of emerging technology [4, 5]. Other uses of elicitation have included conceptual [6] and existing [7] technologies.

This study design was introduced by Wobbrock et al. in 2005 [8] and later popularized by the same team [9]. 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 [8]. The premise around achieving the goal of discoverable inputs stems from distributed cognition [10]. By observing users interact with a system in which the gulf of execution (e.g., barriers of execution) is removed, that user's natural behaviors and interactions can be captured. While these interactions will vary from user to user, an aggregation of multiple users' interactions can be used to derive a consensus set of proposals. This set represents the most common interactions of novice users within this system and experimental setup.

This use of participatory design to derive a consensus set is often a major goal when using elicitation methodologies [11, 12]. That said, observational data is a rich source of intuitions concerning user behavior. Discoveries beyond a consensus set can emerge through the interpretation of that data. These discoveries have included the impact of scale on interaction generation [13, 14], the timing information around co-occurring gesture and speech inputs [1, 15, 16], user modality preference when multiple modality options are available [17–19], or that users prefer multimodal interactions more as task cognitive load increases [20].

The popularity of elicitation studies is evident through the variety of domains that have used them. Some of these domains include multi-touch surfaces [12, 21], mobile devices [22], mid-air gestures [1, 5, 23], television browsing [18, 19], computer-aided design [7], and internet of things home sets ups [24]. Building on the original methodology, researchers have devised alternatives that extend beyond surface-computing devices, such as using multi-touch and mid-air devices in tandem [25,26] or using multi-touch devices to control physical objects through virtual representations of said entities [27]. Imposing constraints on the users' motion has also led to new elicitation studies primarily concerned with defining and investigating gesture sets suitable for both impaired and non-impaired users [28,29].

These studies constitute a massive body of still growing literature spread across many domains of use and disciplines. More than 216 elicitation studies have been run, these include $5,458$ participants, and $3,625$ commands (referents) tested [3]. Alongside the widespread use of this 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" [8,30]. Other changes to the calculation of consensus include between groups metrics [30], production agreement [31], dissimilarity of proposals metrics [31], the addition of speech proposal consensus metrics [18], and statistics to help verify the prevalence of chance agreement [32]. Some studies directly emulate the work of Wobbrock et al. [3,9], while others radically alter the process [33]. There have been variations of the Wizard of Oz systems used [34], the presentation of referents [3], and even attempts to deliberately prime users with a certain mindset [6,29] or mental frame [33].

Elicitation has seen most of its use as a way to derive inputs for a single modality of input. Often this is some form of gesture-based input, either full body or limited to a specific region such as a hand [3]. As new technologies continue to emerge, elicitation is starting to see a divergence from the unimodal standard to use for multimodal input derivation. Examples of this are most commonly for gesture and speech [1,7,15,18,19]. An area that has been unquestioned in the literature today is, "how well does this unimodal interaction design technique transfer into multimodal space?"

Herein lies a concerning facet in this ever-evolving body of literature; there is a scarcity of work examining the results of multimodal elicitation studies and of work examining the reproducibility of elicitation studies [3]. This paper presents a comparison of two multimodal gesture and speech elicitation studies done with basic object manipulations in optical see-through augmented reality (AR) environments [1, 15]. The difference between these studies is the display of referents.

Referents display is often not the main focus of elicitation studies yet, humans possess powerful mimetic skills. Imitation of visible action is considered an inborn skill [35] and imitation of speech plays a formative role in human development [36]. Referents are a means to map commands to input proposals. We find that the choice of referent display is also the choice of how to bias referent imitation in proposals. The examined works were conducted by the same lab with different subjects from the same participant pool [1,15], but did not have equivalent results, thus giving evidence that minor changes in the methodology used in elicitation studies can cause differences in the results obtained.

This work uncovers evidence of a biasing effect inherent in elicitation methodologies, a fundamental concern that has been overlooked thus far. This biasing is caused by participants imitating referents as they were displayed. This concern is grounded in a comparison of data obtained from two studies, two prior replication/reproduction studies, and the literature surrounding human psychology. The comparisons address the differences in elicited proposals across three modalities of inputs: gesture alone, speech alone, and co-occurring gesture+speech. To continue the improvement of elicitation methodology, we propose design recommendations to limit or remove imitation bias.

A consensus set of gestures is a separate contribution of this work. That set of gestures is derived from one of the elicitation studies run during this research. This consensus set is presented with comparisons against the gestures produced in prior mid-air gesture elicitation studies.

1.1 Background

The following section outlines the key terms and processes used in elicitation studies using the work done by Wobbrock et al. in 2009 as an example [9]. This background will allow for a better understanding of the comparisons and conclusions made here.

1.1.1 Terms

- **Elicitation**: prompting potential end-users of a system to generate inputs for that system
- **Agreement**: a measure of how many participants proposed the same interaction
- **Consensus set**: a set of proposed interactions that are highly agreed upon by participants
- **Proposal (sign)**: the input suggested by a user for a given command (referent)
- **Referent**: the command/action which the input proposal will execute
- **Wizard-of-Oz (WoZ)**: a study design where a system's recognition capabilities are emulated by an experimenter
- **Think-aloud**: when participants are asked to describe what lead to the formation of their input proposal
- **Binning**: partitioning proposals into equivalence classes based on pre-defined metrics (i.e. the number of fingers used, hand posture, motion)

1.1.2 Elicitation Protocol

Most elicitation studies follow the same protocol. Commonly around 25 (Median = 20 Standard Deviation (SD) = 4) participants are recruited [3]. 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 and the participant produces an input proposal they think is appropriate using the input modality requested. 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 [9]. This allows users to feel like they are interacting with a live system. In the 2009 study done by Wobbrock et al., 1080 gesture proposals were made by 20 participants proposing both one and two-handed gestures across 27 referents [9]. Referents are commonly specific to one domain or application. The domain chosen in Wobbrock et al.'s work was surface computing and as such referents included *move a little*, *move a lot*, *pan* [9].

Data is most commonly collected through video recordings [3, 37] however, sometimes other means such as skeletal tracking used [19, 31]. The video data is paired with the observational data from interviews and information from the participants gathered by using a think-aloud protocol. Wobbrock et al. used video paired with think-aloud data [9]. The video data is hand-annotated by one or more raters and broken into gestures proposed [1, 9, 15, 38]. These will be very granular gestures with notes on features including the number of fingers used, hand position, and direction of movement [5, 15]. These gestures are binned into equivalence classes based on predefined similarity features or insights from previous work. One such insight is that participants often don't recall or care much about the count of fingers used in a gesture allowing for groupings of one and two-finger similar movement gestures together [6, 9]. Wobbrock et al. used four binning dimensions based on the movement of the gesture [9], other dimensions could be a proposal's semantic features [39]. When skeletal data is collected computer vision techniques can be used to bin gestures eliminating the need for hand annotation and potentially the human bias arising from hand annotation [31]. The binning of proposals is an important step towards removing the individual-level characteristics of the proposals in favor of a more generalizable consensus set.

Agreement metrics are used to quantify consensus across participants on the binned gesture proposals by referent. Often this agreement is measured using the *Agreement Rate* formula, which is a measure of pairs of participants in agreement over all possible pairs [9]. Other metrics such as machine learning techniques [31] and metrics designed for speech [18] exist. Based on these metrics, a set of consensus gestures is proposed for referents that achieve higher than a predetermined level of agreement. This level is often around 0.3 for sample sizes of 20 based on the distributions of agreement from varying participant counts [30]. Design guidelines informed by the proposal space and observational data are a secondary contribution of elicitation studies. The results of Wobbrock et al.'s study was a consensus set of user-defined multi-touch gestures as well as a taxonomy of gesture use [9].

1.1.3 Elicitation Criticisms

Elicitation studies have received criticism in a few areas. First is the impact of legacy bias on input proposals [40]. Legacy bias is when a gesture proposal is heavily informed by participants' interactions with prior technology. An example is a participant saying "F5" when suggesting a speech input for refreshing a browser page [18]. This bias could be leveraged and is not always considered a negative quality of input proposals [23, 41]. Using legacy or near legacy interactions can lead to a more discoverable interaction set as it can mirror users' preconceived mental models of interactions [41, 42]. Several methods exist for reducing legacy bias, further widening the variances found in elicitation procedures. These reduction methods include production (asking for more than one proposal per referents), pairing (grouping participants), and priming (influencing a participant's mindset before eliciting proposals) [40].

Another elicitation concern is the issue of chance agreement which occurs when the input proposal space is small enough that high agreement rates could conceivably be caused by random chance because the agreement rate formula assumes an infinite space of gesture proposals where in actuality that space limited [32]. Tsandilas (2018) suggests that participants will actually cluster around a subset of gestures making the actual proposal space sampled from much more limited [32]. A way to resolve this is to calculate the Fleiss' Kappa coefficient and the associated chance agreement term to assess the impact of chance agreement [32].

1.1.4 Imitation

This paper raises the issue of imitation as a concern needing to be addressed in elicitation methodologies. 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 [35] or learned via self-observation [43] and reinforced from a young age [44]. Regardless of where it arises from, the existence of automatic imitation is the same. Non-human representations (i.e., a wooden hand) stimulate lower imitation than human representations [45]; however, geometric objects seem unaffected by this if their action can reasonably be completed by a human [46], as is the case for referents.

During imitation, perception and action are tightly coupled by a direct perceptual-motor mapping [47, 48]. That mapping connects visual information to proprioceptive information [49]. This mapping is supported by work done on mirror-neurons in primates' pre-motor area, which fire the same way both when acting and viewing an action [50]. Similar neuron activation has been observed in humans [51].

Imitation causes increased activity in the brain's Broca's area which is thought to be involved in speech production [52]. Speech imitation is a skill used from a young age to facilitate language learning [36]. Imitation of speech is debated to either cause erroneous mirror-like activation [53], or to be more difficult to execute due to the muscle groups involved and complexities of language [54]. Under either theory, imitation of action via gesturing is likely antecedent to speech [55].

1.1.5 Referent Display

The goal when presenting a referent is to establish the command to be completed by the input proposal. If eliciting commands for television-based web browsing then a referent would be *refresh page* [18]. *Refresh page* could be presented as text reading "refresh page", an animation of a web page being refreshed, or an experimenter reading the referent aloud. In the case of Morris, 2012, and Nebeling et al., it was both showing the effect of the referent (the animation) and stating its name aloud [18, 19]. Note that both of these studies used gesture and speech as input modalities. The effect of speech imitation can be seen in their results; however, it was never mentioned that reading the referents out-loud contributed to the high overlap between spoken referents and participant speech proposals.

Referents have been presented to participants in a variety of ways which becomes problematic when the elicited proposals may be highly impacted by the choice of referent presentation. Referent display techniques have included animations paired with spoken aloud instructions [18], images [56–59], animations alone [7, 15, 60–63], text alone [1], only read aloud [64], text and animation [38, 65], text and read aloud [23, 29, 66], and the combination of text, reading aloud, and animations [4]. On occasion, the exact form of display is left slightly unclear [6]. With this wide range of prompts used and some evidence suggesting that referents can bias the proposal [3], we believe that further study of the impacts/implications of referent display is merited.

Chapter 2

Related Work

While this work talks about replication, it does not identify the replication of results as a main goal. Comparisons of prior replications in elicitation are used to highlight the differences in results and in particular the differences in results when they can be reasonably explained by the choice of referent display. This section also covers the limited prior work on AR-HMD mid-air gesture elicitation to further motivate the produced gesture consensus set.

2.1 Elicitation Study Replications

The singular reproduction study found in a recent review of gesture elicitation studies [3] was the work done by Sukumar et al. (2018) [39] replicating the work of Wolf et al. (1987) [67], and Welbourn et al. (1988) [68]. The study elicited pen and touch-based gestures on a multi-touch surface for use in text editing applications [39]. Sukumar et al. used a modified elicitation methodology based on the work of Wobbrock et al. in 2005 [8, 39].

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 [39] as opposed to pen and paper [67, 68]. That difference was further pronounced by telling participants they were interacting with a live recognition system compared to paper alone causing a difference in perception that could impact the participant's production of proposals [69, 70].

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. note differences in the referents "join", "split", and "new paragraph" which were conceptually similar to the commands used in the previous experiment [68], but had a different wording [39]. They go on to speculate that the differences in results are caused by those variations in terminology, citing other work that used the same terms to produce similar results during multi-touch elicitation [71]. The differences in terminology used are akin to differences in referents used. The legacy biases inherent in the participants caused by 30 years of technological advancement may have also contributed to differences in results [40].

The sole replication study found by that same review of gesture elicitation studies [3] was the study by Nebeling et al. (2014) replicating the work of Morris (2012) [18]. Both of these studies elicited gesture and speech commands for a television-based web browser equipped with a Microsoft Kinect using 25 participants each. Participants were put in pairs with a single triad and asked 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 [18] as closely as possible, omitting only a few of the original referents.

Participants' 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 [18, 19]. 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 to the Microsoft Kinect 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 replication.

Regardless of the causes of the differences, the examination of these two studies brings to light the issue that proposals may not replicate well. More work is needed to examine the potential causes of this failure to replicate. This paper contrasts two studies that were run in the same controlled environment with different participants from the same participant pool, and in the same year, allowing for removal of the concern of differences in time and some of the concern of differences in prior device exposure impacting the elicited proposals between the two studies.

2.2 Gesture Elicitation

Several studies have created gesture sets using gesture elicitation [11,12]. Most of these works focus on gestures for domains outside of mid-air 3-Dimensional use. These include studies on multi-touch devices [12,21], mobile devices [22], internet of things home setup [24]. Commonly these studies will impose constraints on what a user can propose such as asking for pointing gestures [72,73], paddling gestures [74], or 2-dimensional (2D) gestures [73,75].

There has also been some work on gesture elicitation in augmented reality; however, it was done using a VR headset and rendering hands which excluded both the rendered object's opacity and the user's real world view that AR provides [5]. Recent work on mid-air gestures has been done for smart devices/rooms [66,76].

The work presented here is unique in that there are no constraints imposed on input proposals. Participants are free to generate any proposal that they feel is best suited to the referent displayed. A second major difference is that this work was conducted using an optical see-through AR-HMD. The gestures produced within an AR-HMD may vary due to the user's perceived state of the system and the visual feedback of their hands.

Chapter 3

Methods

This paper used the annotated data from the two studies performed during this course of research [1, 15]. 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 [1, 15]. These two experiments were similar apart from the way the referent was displayed. The first study used text referents (top of Figure 3.1), referred to as "E-Text" [1]. The second study used animated referents with no text was shown except the modality to be used (bottom of Figure 3.1), referred to as "E-Animated" [15]. Statistics were run in R version 4.0.2.

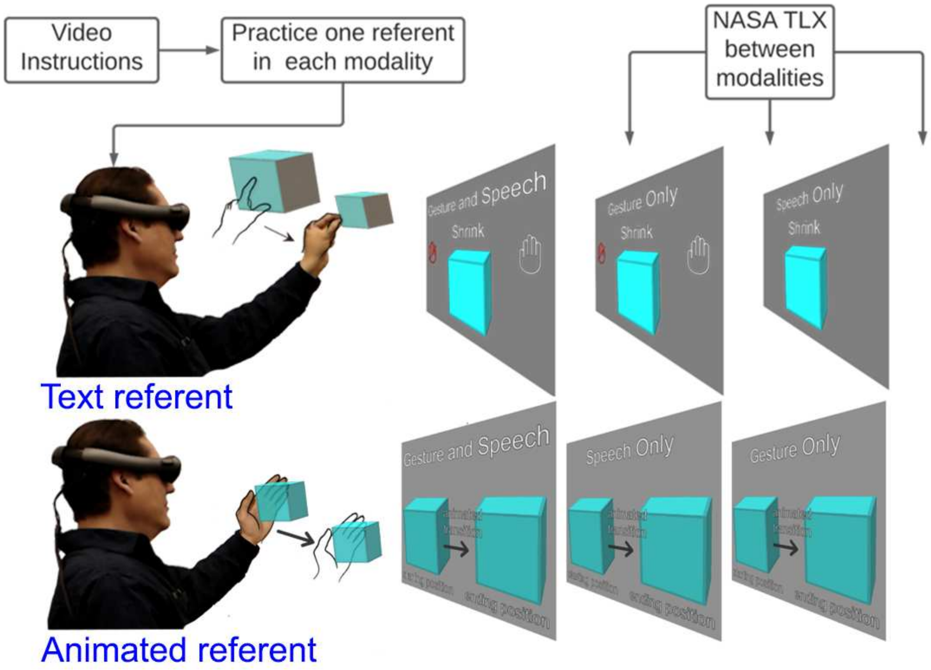

Figure 3.1: High level study flow, Top: text referent (E-Text), Bottom: animated referent (E-Animated)

Both experiments were run on a Magic Leap One optical see-through augmented reality head-mounted display. The system for each was developed in Unreal Engine using version 4.23.0 for E-

Text and 4.24.01 for E-Animated. Each experiment was developed on a Windows 10 professional computer with an Intel i9-9900k 3.6GHz processor and an Nvidia RTX 2080Ti graphics card. The internal camera on the Magic Leap One, an external 4k camera, and a head-mounted GoPro Hero 7 camera each recorded video.

3.1 Study Design

The two experiments examined here each used a Wizard-of-Oz (WoZ) design. Participants were videotaped from both ego-centric and exo-centric viewpoints 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. If the modality was speech then any utterance proposed was accepted causing the experimenter to trigger the system's recognition of 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) and Virtual Reality (VR) and AR environments. 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. The video informed participants that they would be asked to complete a series of object manipulations using different modalities of input and within each modality any input they proposed was acceptable. Participants were given a practice round where they generated a proposal for a color change referent in each modality. During the practice block for E-Text, participants could test the system's on-screen hand detection which would alternate between showing a red hand with a line through it or a white hand indicating that their hand was either in the tracking range of the device or not (top Figure 3.1). During the practice for either experiment, the participant could ask questions about the experiment and what was expected of them.

After the practice round, participants were shown interaction modalities in a counterbalanced order. Within each modality, the referents were shown in random order. For example, participants may have seen “gesture and speech” first, then after generating proposals for each referent, see the next modality condition (gesture alone or speech alone in this example). In each trial the participant was shown a cube rendered approximately 50 cm away from them, centered in their viewport. 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 condition’s perceived workload [77]. The NASA-TLX is a survey used to rank participants’ perceived workload across six subcategories conditions; mental demand, physical demand, temporal demand, performance, effort, and frustration [77]. The scores from the subcategories are combined to give an overall score.

Some objects are likely to prime users to form specific hand-shapes when interacting with them. This priming is normally caused by the affordances of the object. For example, a coffee cup will likely prime gestures to be shaped to match the handle of a coffee cup where a plate might cause more gestures that mirror handling a plate. This work chose a cube as the object to be interacted with. A cube represents a simple object that can help remove some of the impacts of object affordances on proposals [64]. While this choice limited some priming for specific object grip gestures (i.e., grabbing a handle of a cup) it did cause some proposals to have flat-handed gestures emulating physical contact with the surface of the virtual cube. In E-Animated it allowed for visual cues of rotations not seen in with a cylinder or sphere.

3.1.1 Differences in Methods

In E-Text, the referents were shown as text and read aloud (top of Figure 3.1) [1]. 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 [15]. Participants would see a blue screen, then the rendered cube and modality information (right of Figure 3.1). After a two-second delay, the referent would execute the same animations shown in E-Text (with exceptions to the abstract referents). 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 had used to generate that input proposal creating the belief in participants that the system was live but disabled for them. This design choice was made to try to capture feelings of interaction with a live system as seen in E-Text.

With either design, differences were expected based on the level of priming caused by either the animations or the text [1, 15]. When the referents were shown as text, the proposed speech closely follows that text. The use of text referents in elicitation is common [4,23,29,64,66]. When prompting the user with animations, the gestures produced are more likely to be primed by the movements of the objects. This design is also common within elicitation studies [7,15,60–63]. In the few multimodal elicitation studies that have been run, the impact of these decisions has never been stated [18, 19]. This omission has indirectly implied that common elicitation methodologies may be equally valid when dealing with multimodal or non-gesture inputs.

3.1.2 Referents

Table 3.1: Referents used by category

Translation	Rotation	Abstract	Scale
Move (Left / Right)	Roll (C / CC)	Create	Enlarge
Move (Up / Down)	Yaw (Left / Right)	Destroy	Shrink
Move (Towards / Away) from self	Pitch (Up / Down)	Select	

Legend: C: Clockwise; CC: Counter Clockwise

These studies used referents (i.e., commands) that are considered canonical manipulations for 3D user interfaces [78, 79]. The canonical referents used were selection, scaling, translation on each axis, and rotation about each axis. The abstract referents of create, and destroy were also included to increase the usability of these results in generic interactive environments. These referents can be viewed by category of action in Table 3.1. For referents that were not *select*, participants were told they could assume objects were already selected. These referents were selected to generate a set of user-derived interaction techniques that would be usable in generic AR building tasks, an example task being the construction of a virtual house in an AR environment using virtual Lego-like blocks.

3.1.3 Pilot Studies

Before running these experiments, one survey, two pilot studies, and one observational session were administered to different groups of participants. The pilot survey was administered to an entry-level computer science course for students with non-traditional backgrounds (N=35). This survey asked participants to define the referents that were used in the two elicitation studies (Table 3.1).

Two versions of the main elicitation experiment were run on pilot groups consisting of 6 people each. In one, referents were displayed as text (top of Figure 3.1); in the other, the action of the referent was shown as an animation. As an example, if the referent was *move left*, in the first set up the screen read "move left" and participants were asked to propose a command to execute that referent (similar to [9, 18]). Upon generation of that proposal, the virtual object would move. In the second design, the virtual object would move before participants were asked to generate an appropriate command proposal (similar to [7]).

An observational session was run where 5 participants were shown different animations for the referents *create*, *delete*, and *select*. After seeing each animation those participants were asked to state what the animation was showing (i.e., moving left, selection). Variations of animations for *create*, and *delete* were shown using no animation and using a slow materialization where the object was loaded or removed over time with particle effects. Animations shown for *select* included arrows pointing at an object, an object bouncing, and an object being highlighted. This was done to help solidify the animation choices for the abstract referents *delete*, *create*, and *select*. Raw counts were used when interpreting the pilot study data.

3.1.4 Participants

The pilot survey was administered to 35 incoming computer science undergraduate students. The pilot studies were run with 6 participants each using the same recruitment methods as were used during the final experiments. The observational session had 5 participants volunteer from the same pool of participants recruited for the final studies.

Each study consisted of 24 volunteers (E-Text: 4 female, 20 male; E-Animated: 10 Female, 14 Male). 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 Microsoft HoloLens 1 usage before this experiment. In E-Animated five participants reported weekly use of VR. Only two of those participants used VR more than 5 hours weekly (5 hours, 10 hours). The other three participants reported 1-3 hours of VR use weekly. 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.

The sample size of 24 participants per study was grounded in prior work. Most elicitation studies use a median of 20 participants with a mean of 25 and a standard deviation of 4 [3]. Additionally, when agreement rate was suggested as a metric for elicitation study, a sample-size of 20 was referenced as appropriate [30]. Given that the conditions of input modality presentation were counterbalanced, sample sizes of 18 and 24 were considered, 24 was chosen as the most appropriate count. No participant was able to volunteer for or take part in more than one study or survey.

3.2 Data Preparation

After the experiment, each participant had data for the demographics survey, as well as 3 video streams and a NASA TLX survey response for each of the 3 input modality conditions. This totaled 9 video segments, 1 demographics questionnaire results, and 3 NASA TLX survey results per participant. The 3 video streams captured were from the ego-centric head-mounted go-pro camera, the magic leap 1's ego-centric camera, and the exo-centric video camera.

The ego-centric Go-Pro camera videos were hand-annotated to produce the data that was interpreted during these studies. The exo-centric camera was used as a fallback if the ego-centric camera video was unusable for a given referent (e.g., a user's hands were out of the frame). 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. A set of videos for a single participant would have a total of 51 input proposals broken down into 17 gesture-only proposals, 17 speech-only proposals, and 17 gesture+speech proposals (e.g., a gesture proposal with a co-occurring speech proposal given for a single referent).

3.2.1 Pilot and Survey Data

Demographics characteristic information was collected and merged into a single file for each experiment. This data included prior device use information, age, gender, eye-sight, and major or job type information. The NASA TLX data were merged into a single file per input modality condition within each study. At the end of this process, each study would have one demographics information file, and three NASA TLX files, one for each input modality condition. Each NASA TLX file has the data from all participant's responses for its corresponding input condition. Raw scores, averages, medians, and standard deviations were used when analyzing the NASA TLX and demographics survey data.

The pilot survey data were reduced over all participants such that each participant would have 1 response to each referent that was either correctly or incorrectly defined. The data from the pilot studies were identical to the raw video data collected during the full experiments. For E-

Text the speech proposals were marked as either repeating the referent or not repeating it. The gesture proposals and time information were not annotated from those videos. Instead, these videos were watched by the experimenter to gain a better understanding of how much text or animation biased participant input proposals. Experimenters went through the videos of the observational session and noted which animations were incorrectly identified and which were correctly identified. This resulted in a list of referent animations and counts of when they were or were not identified correctly, and a short text field indicating what participants identified the animation as if it was incorrect. Raw counts and averages were used when interpreting the pilot survey, pilot studies, and observational session data.

3.2.2 Gesture Data Preparation

Gestures were annotated from the video at a granular level then binned into high-level equivalence classes. 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 on the y-axis. These equivalence classes are reasonable given that users care less about the count of fingers used than the hand pose used [26]. This resulted in each participant having a binned identification (ID) number for each gesture proposed. These were recorded per referent and modality condition such that a participant would have 17 gesture proposals for gesture alone and 17 gesture proposals from the gesture+speech condition. Agreement metrics were computed using these gesture IDs.

3.2.3 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” 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. This resulted in a participant having 17 speech only proposals and 17 speech proposals for co-occurring gesture+speech interactions. These proposals were matched to their corresponding referents and input modality conditions. These binned utterances and syntax choices were used when computing the speech consensus metrics.

3.2.4 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.

For each referent and proposal, the experimenters hand-coded the timestamps from the video for when the gesture portion of the proposal and the speech portion of the proposal were initiated. Instances where a hand moved and then immediately returned to the rest position before executing an actual gesture were excluded. Similarly, for speech proposal instances where participants said “um” or another filler word with a pause before a second utterance with contextual meaning were excluded. The start time for the gesture portion of the interaction was subtracted from the start time of the speech portion of the interaction. This gives a time coding where time 0 was always gesture initiation and the time listed represents the time delay between that gesture and its corresponding speech proposals. These times could be negative, zero, or positive. A negative time occurs when speech is initiated before gesture and a positive time when speech was initiated after the gesture. This resulted in each participant having 17 time values, one for each referent during the gesture+speech input condition. These time values were averaged for each participant before analysis giving 24 time values for each study (one for each participant).

3.3 Analysis Performed

This section will cover the analysis and higher level goals used in interpretation of the data from these experiments.

3.3.1 Pilot Data

The pilot survey data was analyzed using the raw counts of the correctly and incorrectly answered responses per referent across all surveyed participants. The pilot studies were analyzed by experimenters watching the video of the sessions and by using the raw counts of times the referent was repeated in the speech and gesture+speech conditions of E-Text. The raw counts of participants that correctly or incorrectly identified the animation shown during the last observational session were used to help inform animation design for the elicitation experiments.

3.3.2 Gesture Metrics

For gesture analysis, the main metric used was Agreement Rate ($\mathcal{AR}$). The formula for $\mathcal{AR}$ is shown in Equation 3.1. $\mathcal{AR}$ is a measure of how much participant agreement there is for a given referent. Given the sample size used (24) a $\mathcal{AR}$ of .3 is considered high agreement, meaning if the referent *select* achieved an $\mathcal{AR}$ of .5 then the most frequent proposal for *select* would be considered highly agreed upon and thus discoverable to novice users of this system [30]. In Equation 3.1, P is the set of all proposals for referent r, and P_i are the subsets of equivalent proposals from P [30].

$$\mathcal{AR}(r) = \frac{\sum\limits_{P_i \subseteq P} \binom{|P_i|}{2}}{\binom{|P|}{2}} \tag{3.1}$$

The rate of individual gesture proposals were compared across referents between the two studies. An example would be saying that the gesture with the binned ID “08” was proposed 9 time when prompted with animations and 2 times when prompted with text. Note that within each referent and input condition a participant can only have a single gesture proposal. This analysis aimed to outline the differences in proposal formation and frequency as dependent on referent display.

3.3.3 Fleiss' Kappa

To calculate the level of chance agreement (P_e) within the elicited proposals the chance agreement term (Equation 3.2) is used. This term stems from the calculation of Fleiss' Kappa [32]. In Equation 3.2 m is the total number of proposals, n_{ik} is the number of participants proposing proposal i in bin k, n_i is the total number of proposals for proposal i. The term π_k reflects the chance that a rater classifies an item into category k based on the times that category has been used across the data. q is the space of possible proposals. The $\mathcal{AR}$ can be inflated by chance agreement if the total number of distinct gestures proposed during the study is low. The use of P_e allows us to compare the $\mathcal{AR}$ value with the level of chance agreement to determine if the $\mathcal{AR}$ is inflated because of high levels of chance agreement.

$$p_e = \sum_{k=1}^{q\_} \pi_k^2, \quad \pi_k = \frac{1}{m} \sum_{i=1}^{m} \frac{n_{ik}}{n_i} \tag{3.2}$$

3.3.4 Speech Analysis

Speech was analyzed using two metrics of agreement. The first is max-consensus ($\mathcal{MC}$). $\mathcal{MC}$ is the percent of participants proposing the most common utterance proposal [18]. If 12 participants proposed the utterance "move left" for the referent *move left* and 5 propose "left", 2 propose "move", and 1 participant proposes "sideways" the $\mathcal{MC}$ equals 60%. The second speech metric is the consensus-distinct ratio ($\mathcal{CDR}$). $\mathcal{CDR}$ is the percent of proposals for a referent that has over a baseline of 1 participants proposing them [18]. In the above-mentioned proposal scenario, the $\mathcal{CDR}$ is 75%. $\mathcal{MC}$ and $\mathcal{CDR}$ were averaged across referents to gauge the general level of difference in metrics between the two studies.

These metrics capture the peak and spread of the speech proposal space [18]. If a referent has a proposal with a high $\mathcal{MC}$, that proposal is considered discoverable to novice users of this system. Alternatively, a high $\mathcal{CDR}$ means that a referent has a high amount of disagreement on the best choice of proposals for that referent between participants. These are not exclusive metrics. It is possible to have a referent with a single highly proposed interaction (i.e., a high $\mathcal{MC}$) and a number of proposals that are suggested by single participants (i.e., high $\mathcal{CDR}$). This would imply that there is a clear most common utterance but not a clear second place or alternative choice utterance. By comparing these metrics and the top choice utterances for the speech proposals from the speech-alone and the gesture+speech input conditions, the differences in the elicited speech proposals across the two choices of referent display can be assessed.

Speech was also analyzed using each of the binned syntax's rate of use as a percentage of all syntax use. The goal of the syntax analysis was to outline the impact of referent display on speech proposal syntax.

3.3.5 Time Windows Analysis

The time window information as annotated from the videos of the participants interactions during the study were first analyzed using Shapiro-Wilk tests of normality. Then, as informed by the results of the previous test, Wilcoxon rank sum tests were used to assess the median time between gesture and speech initiation. The goal of the Wilcoxon rank sum test was to find what time windows could be reasonably expected, if there were differences between the median values of these time windows as caused by referent display, and to visually compare the differences in time windows between the two experiments.

3.3.6 NASA TLX Analysis

Means and standard deviations were computed for NASA TLX results. The differences in NASA TLX results between input modality conditions were compared using Welch Two Sample T-Tests. Sharpiro-Wilk tests for normality were run first to see which category of t-test would be appropriate.

3.3.7 Consensus Set

The gesture consensus set is derived using most frequent gesture proposal for the gesture and the gesture+speech conditions. Recommendations of use are grounded in the $\mathcal{AR}$ results for referents in the gesture and the gesture+speech conditions.

Chapter 4

Results

The goal of this work is to assess the impact of referent display on the resulting input proposals during elicitation studies. This paper also provides a consensus set of gestures from E-Text. As elicitation continues to gain popularity in assessing multimodal input design, this meta-examination is critical to the ongoing improvement of elicitation methodology. This study represents the first comparison of elicitation studies that were done by the same team, with the same subject pool, in the same year. This examination can provide insight into how minor changes in design can impact results.

The results of the original studies were divided into four comparisons: gesture versus gestures from gesture+speech, speech versus speech from gestures+speech, gesture+speech alone, and surveys or additional data. This paper will examine the differences in these studies in a similar manner comparing first the impact of showing a participant an animated referent compared to a text-based referent on gesture, then speech, then the combination of gesture+speech, and then a comparison of the survey results. The results section concludes with a consensus set of gestures for use with generic object manipulations in AR.

4.1 Pilots and Observation Studies

4.1.1 Pilot Survey

The definitions for the translation, scale, and abstract referents were each correct for more than 30 of the responses of the pilot survey. The rotational referents, particularly pitch and yaw, had 2 of 35 people correctly define them. Due to these results, a visual explanation of rotation commands was added to the video instructions using a stuffed animal to minimize biasing the subjects. In the past, using an airplane metaphor (palm down with fingertips as the nose) biased the subjects to perform "airplane" gestures [80].

4.1.2 Observation Session

During the observational session using varied animations for the abstract referents, 3 participants accurately identified the animated referent create or delete where no participants identified the unanimated referent correctly. *Select* was more difficult to animate. The highlighting condition had 2 participants correctly identify the animation as a selection. No participants correctly identified selection in the other animations. Based on this, the results for the abstract referents in E-Animated were expected to be highly biased by the animation used. Additionally, the highlight selection and particle effects creation/deletion animations were chosen for use.

4.1.3 Pilot Studies

During the speech block of the pilot study where referents were displayed as text, participants would commonly repeat the referent displayed. For example, if the referent was *move left* the utterance was also “move left”. In the pilot study without text, for simple translations, the most frequent utterances were “move” and the direction such as “left”. For the referents that were not translations, the repetition varied more between the two conditions. For the rotational referents, most participants exactly mirrored the displayed text. These participants were from the same pool of students that the pilot survey was administered to indicating that this repetition occurred even when the likelihood of participants being familiar with the rotation terms was low. For the animation condition, the proposals for rotational referents were most commonly “rotate” with a direction (i.e., left, up).

In the version of the pilot study where referents were shown as movement (bottom of Figure 3.1), people would nearly always propose a gesture that used a motion that very closely matched the object’s animated motion. For rotations, people would twist their wrists into uncomfortable positions to try and match the object’s motion. For the abstract referents, people’s gestures would mirror whatever animation was shown. If the virtual object was materializing from right to left, participants’ hand moved from right to left. None of the participants understood what was being asked of them when the referent was *create* and the virtual object appeared with no animation. The effects of referent animations biasing gesture production can be seen in the elicited gestures of prior work [7]. Examples include the proposed gestures for the *orbit* and *pan* referents which have participants' top choice gestures mirroring the visual motion of those referents.

4.2 Gesture Comparisons

4.2.1 Agreement Rate Comparisons

Table 4.1: Agreement rates per referent compared across E-Text and E-Animated with absolute differences shown

	Create	Delete	Enlarge	Move Away	Move Down	Move Left	Move Right	Move Towards	Move Up	Pitch Down	Pitch Up	Roll C	Roll CC	Select	Shrink	Yaw Left	Yaw Right
Gesture																	
E-Text	0.08	0.11	0.29	0.55	0.37	0.48	0.47	0.47	0.34	0.17	0.14	0.54	0.66	0.84	0.24	0.32	0.27
E-Animated	0.21	0.08	0.31	0.35	0.34	0.51	0.44	0.28	0.49	0.14	0.29	0.51	0.34	0.10	0.13	0.19	0.19
Difference	0.12	0.02	0.02	0.20	0.02	0.03	0.04	0.19	0.14	0.02	0.15	0.03	0.32	0.74	0.11	0.13	0.08
Gesture+speech																	
E-Text	0.06	0.10	0.28	0.47	0.46	0.62	0.47	0.34	0.38	0.24	0.23	0.48	0.35	0.82	0.29	0.31	0.30
E-Animated	0.13	0.08	0.28	0.56	0.30	0.74	0.69	0.38	0.40	0.27	0.34	0.32	0.59	0.10	0.20	0.26	0.33
Difference	0.07	0.02	0.00	0.08	0.17	0.12	0.22	0.04	0.02	0.03	0.10	0.16	0.24	0.72	0.09	0.05	0.03

Legend: C: clockwise, CC: counterclockwise, referents and differences with high levels of agreement rates are highlighted, differences are absolute values

The agreement rate ($\mathcal{AR}$) metric is a measure of the consensus of participants' proposals for a given referent [30]. While these rates are typically not to be compared across studies due to the potential for the number of participants to impact the calculation of $\mathcal{AR}$, the equivalencies in design between the two experiments make it reasonable to compare the $\mathcal{AR}$ here. The agreement rates for the two studies can be seen in Table 4.1. For 9 out of 17 referents in the gesture condition and 7 out of 17 referents in the gesture+speech condition, the difference in $\mathcal{AR}$ is below $.1$. As a reference point, given the studies sample size and the correlated Fleiss' Kappa [30, 32] (E-Text: $.057$, E-Animated: $.054$) [1, 15], an $\mathcal{AR}$ of 0.1 to 0.3 is considered medium agreement, and a rate of 0.3 to 0.5 is considered high agreement. This indicates that the differences in referent display between the two experiments caused a medium difference in $\mathcal{AR}$. For 4 out of 11 referents in the gesture condition and 3 out of 12 referents in the gesture+speech condition, the agreement changed from being considered high agreement to medium agreement. This means that 30% of the referents with high agreement for one referent display had medium agreement when the alternative display was used.

Stark differences in the agreement are found in the *select* referent which required an abstract animation (Table 4.1). The difference in $\mathcal{AR}$ for the *select* referent between experiments was 0.74 for the gesture condition and 0.72 in the gesture+speech condition. In E-Text, most participants tapped the cube with a single finger causing *select* to have the highest $\mathcal{AR}$ out of all of the referents. In E-Animated, proposals became far less consistent due to varied interpretations of the referent's meaning. This difference is expected to be caused by participants misinterpreting the animation for *select* which was the cube becoming "highlighted" by increasing its glow and hue. The only other referent that had a high level of difference in $\mathcal{AR}$ between the two studies was the *Roll Counterclockwise* referent in the gesture only condition. Roll counterclockwise had a difference in $\mathcal{AR}$ of 0.32.

Both the referent *create* and *destroy* had low $\mathcal{AR}$ in either study which in turn made the differences in $\mathcal{AR}$ between the two referent displays limited. There was an increase in $\mathcal{AR}$ of 0.12 for the referent *create* when transitioning from E-Text to E-Animated. This may indicate that the animation used for *create* increased participants' consensus on which gesture proposal was best suited for it.

4.2.2 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 proposals from the gesture alone condition are shown in Figures 4.1, 4.2 and, 4.3. The heat-maps for the gesture proposals from the gesture+speech condition are provided in Appendix A Figures A.1, A.2 and, A.3. These heat-maps do not list any proposals that were elicited once across both experiments. This step reduced the visual clutter shown in these heat-maps by reducing the number of rows (gesture proposals) displayed. As these cases are removed, column totals may not sum up to 24. The heat-maps from both the gesture alone and the gesture+speech condition are similar in proposal space and differences between the two experiments. This section will use the heat-maps for the gesture alone condition to outline the differences between gestures proposed using text-based referents and animated referents.

In these heat-maps the y-axis provides a short description of the gesture proposals, the x-axis lists the referents across each of the two experiments. The individual cells in the heat-maps represent the frequency of proposals for a given gesture with darker cells representing increased proposal frequency. As an example, the first two columns of Figure 4.1 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 (Figure 4.1) and the rotation referents (Figure 4.2) are often quite similar. The difference for most referents is a slightly increased variety of gestures proposed as exhibited in the minor increase in the number of distinct proposals in the columns for the referent *pitch down* in Figure 4.1. The scale referents also show similar proposal frequency between the two experiments (far right on Figure 4.3).

The referent *select* has the largest deviation in gesture proposals between experiments (middle pair of columns in Figure 4.3). In E-Text there is one gesture proposal elicited 22 times wherein E-Animated the most frequent proposal slot is tied with 5 participants suggesting each. This discrepancy is likely caused by misinterpretation of the animation for the *select* referent.

The *delete* referent 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 is *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 simple referents are relatively minor. Abstract referents may be more impacted by the shift from text to animated, but 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 differences in proposals for *delete* were limited indicating that this animation is similar to the concept of the word delete. The differences in *select* show that animations can result in a noticeably different proposal space (Figure 4.3).

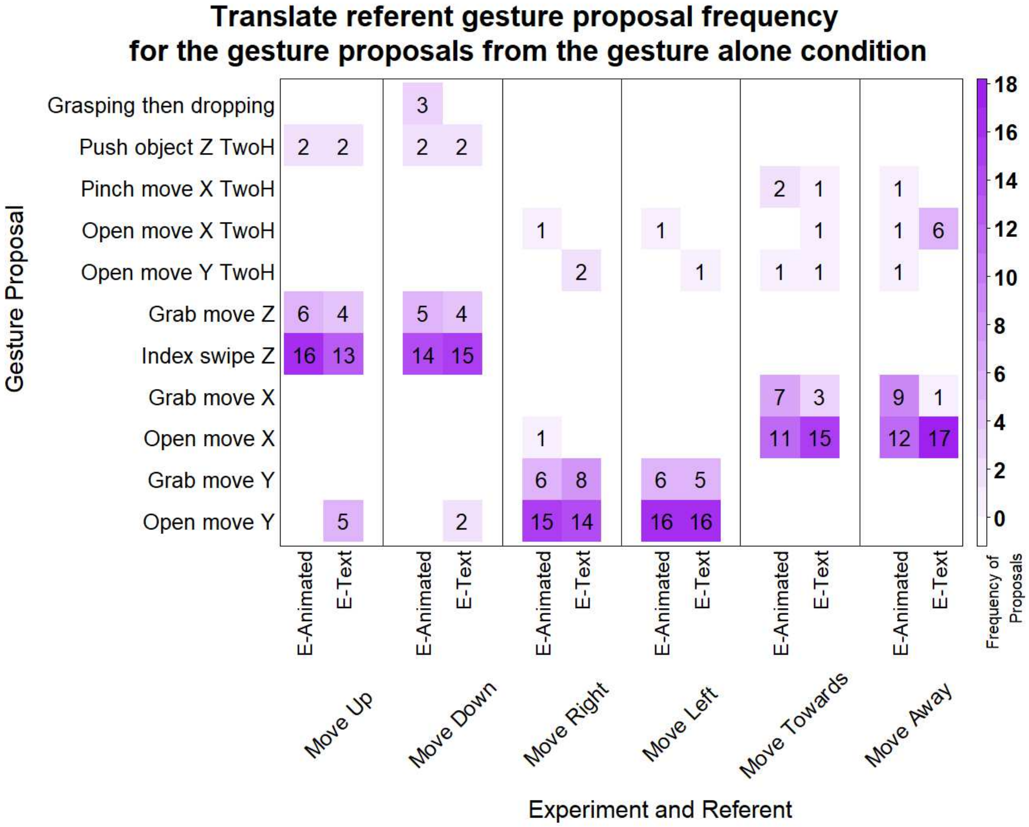

Figure 4.1: Heat-map of common gesture proposals by referent and experiment (translation referents only) **Legend**: Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed

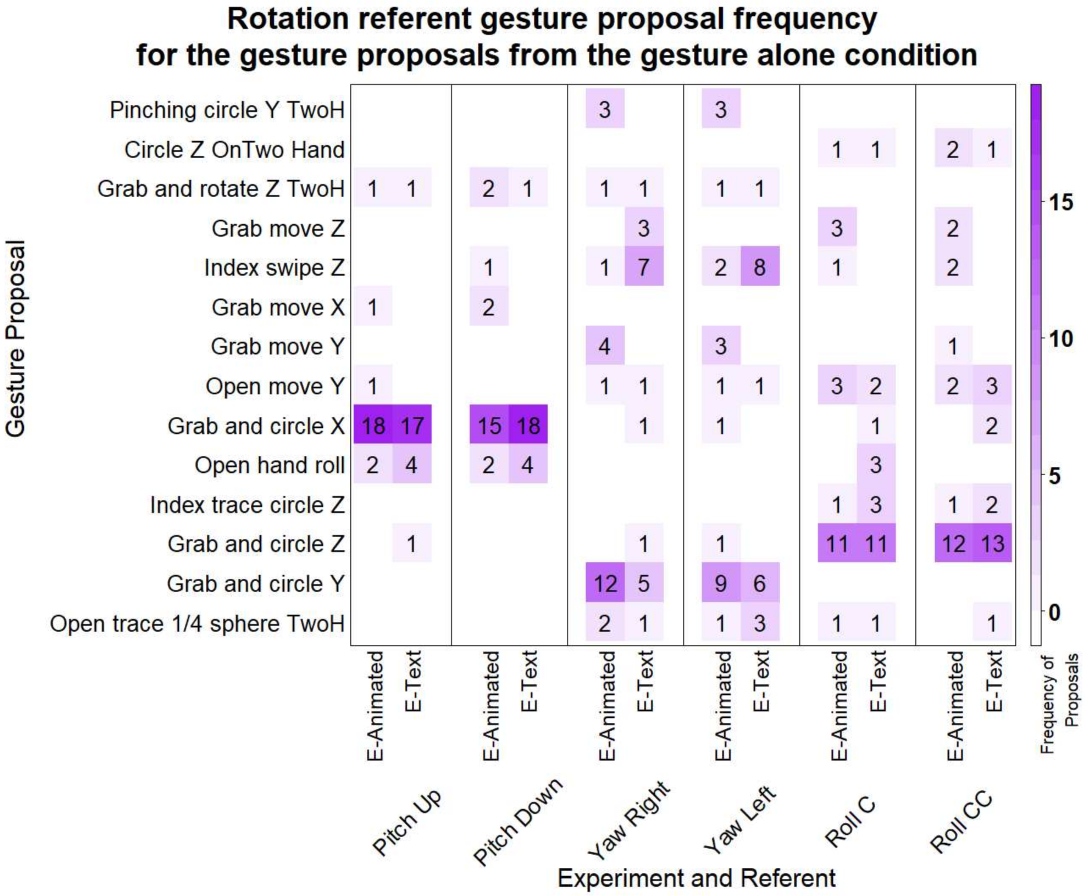

Figure 4.2: Heat-map of common gesture proposals by referent and experiment (rotation referents only) **Legend**: E-T: E-Text, E-A: E-Animated, Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed, CC: counterclockwise, C: clockwise

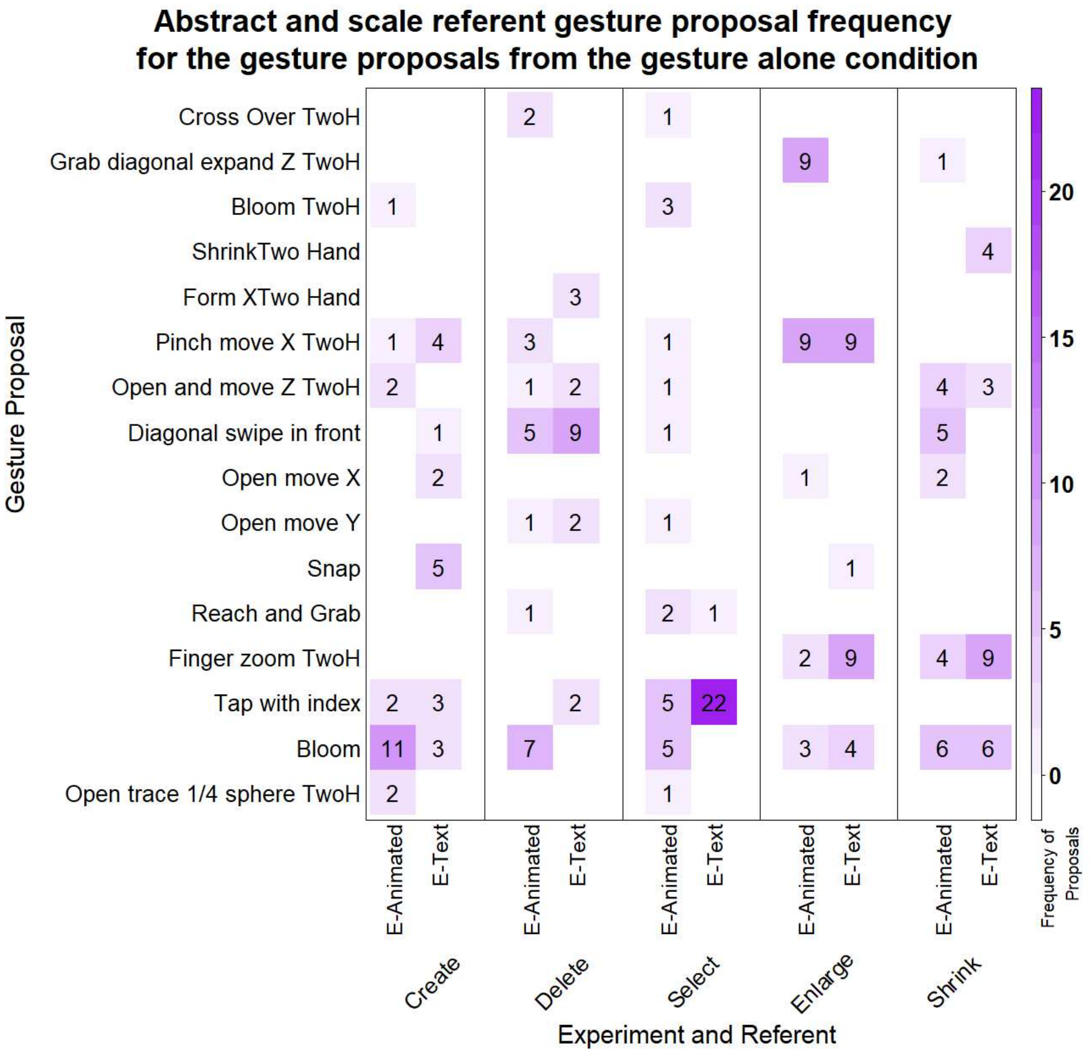

Figure 4.3: Heat-map of common gesture proposals by referent and experiment (abstract and scale referents only)

Legend: E-T: E-Text, E-A: E-Animated, Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed

4.3 Speech Comparisons

4.3.1 Syntax Usage Comparisons

Table 4.2 displays the rates of each syntax type used by condition for each experiment. as well as the differences in syntax used across E-Text and E-Animated. The difference in syntax used during the speech-only condition between the two experiments was less than a 5% change. In the gesture+speech condition, two of the syntax structures used had a shift of more than 10% while the other differences were less than 4%. This implies that the type of syntax used when generating speech proposals was largely unaffected by the referent display choice. The largest observed difference was a shift of nearly 10% syntax use between using only an action phrase and using only a direction phrase. In E-Text direction alone saw 11.76% use and action alone was used 28.19% of the time. In E-Animated there was nearly a 10% shift where action alone was used 38.48% of the time and direction alone was used 1.72% of the time. This is evidence that animated referent display may be more likely to elicit an action phrase where text is more likely to elicit a direction phrase. That said across most categories referent display showed minimal impact with most differences being less than 5%.

Table 4.2: Frequency of syntax used across experiments by condition with absolute differences

Experiment	<action>	<action> <direction>	<action> <object> <direction>	<action> <object>	<direction>	<other>
Speech						
E-Text	24.75%	50.25%	12.75%	5.64%	6.13%	0.49%
E-Animated	28.19%	47.06%	14.22%	9.31%	1.23%	0%
Difference	3.44%	3.19%	1.47%	3.67%	4.9%	0.49%
Gesture+speech						
E-Text	28.43%	43.87%	10.54%	4.41%	11.76%	0.98%
E-Animated	38.48%	39.95%	12.99%	6.86%	1.72%	0%
Difference	10.05%	3.92%	2.45%	2.45%	10.04%	0.98%

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

4.3.2 Speech Proposals and Agreement Metrics Comparisons

The most common proposals in E-Text for both the gesture+speech and speech blocks were always the referent as it was displayed (Figure 4.3). The $\mathcal{MC}$ for those proposals was uncommonly high in every case, likely caused by participants imitating the referent. E-Animated had more variance in the proposal space. In 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 the translation referents.

The largest difference in $\mathcal{MC}$ between the two studies was 66.67% in the speech-alone condition and 58.37% in the gesture+speech condition. The average difference in $\mathcal{MC}$ between studies was 42.45% for speech-alone and 42.15% for gesture+speech. The smallest difference in $\mathcal{MC}$ was 16.67% in the speech condition and 20.83% in the gesture+speech condition. These numbers imply that while in some cases the difference in $\mathcal{MC}$ between referent displays may be lower, for most referents these differences are much larger. E-Text had an average $\mathcal{MC}$ of 75.26% (69.36% in gesture+speech) where E-Animated had an average $\mathcal{MC}$ of 32.81% (27.21% in gesture+speech). These differences suggest 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 reported less than a third of participant agreement. The proposals that repeated referents in E-Text and the differences in $\mathcal{MC}$ between the studies are strong evidence that text primed users' speech proposals are likely to report an inflated $\mathcal{MC}$.

The $\mathcal{CDR}$ between these two studies was also varied. Often the $\mathcal{CDR}$ in E-Text was higher than in E-Animated meaning that E-Text had a more narrow 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 and 0.47 average $\mathcal{CDR}$). Alternatively, in E-Animated where no text was shown, there was a much more varied space of speech proposals (0.42 and 0.39 average $\mathcal{CDR}$). This difference is largest with the *move left* and *move right* referents in the gesture+speech condition which both had a $\mathcal{CDR}$ of 1 in E-Text and had a $\mathcal{CDR}$ of .2 and

.33 respectively in E-Animated. These differences in $\mathcal{CDR}$ are further evidence that text based referents can impact the speech proposals generated during elicitation when compared to animated referents, resulting in a less varied speech proposal space.

The results from E-Animated show more variety in top proposals as well, particularly in the rotational referents where ‘spin” was a common utterance. The abstract referents in E-Animated showed the impacts of priming. For *create* and *delete* the top proposals were “appear” and “disappear” which were similar to the referent but closer to the animation used. *Select* had a top proposal of “change” which is much further from the referent while still close to the animation used.

Table 4.3: Speech proposal comparisons by input condition and experiment with absolute differences and column averages

E-Text			E-Animated			Difference	Difference
Top proposal*	$\mathcal{MC}$	$\mathcal{CDR}$	Top proposal	$\mathcal{MC}$	$\mathcal{CDR}$	$\mathcal{MC}$	$\mathcal{CDR}$
Speech							
create	75%	0.33	appear	41.67%	0.18	33.33%	0.15
delete	91.67%	0.92	disappear	50%	0.57	41.67%	0.35
enlarge	66.67%	0.67	enlarge	37.5%	0.36	29.17%	0.31
move away	54.17%	0.42	move back	25%	0.38	29.17%	0.04
move down	79.17%	0.58	drop	33.33%	0.44	45.84%	0.14
move left	87.5%	0.71	move left	37.5%	0.44	50 %	0.27
move right	87.5%	0.75	move right	41.67%	0.44	45.83%	0.31
move towards	37.5%	0.38	move forward	20.83%	0.36	16.67%	0.02
move up	79.17%	0.67	move up	54.17%	0.33	25%	0.34
pitch down	79.17%	0.79	rotate	20.83%	0.46	58.34%	0.33
pitch up	79.17%	0.75	rotate away	16.67%	0.5	62.5%	0.25
roll C	70.83%	0.62	spin right	20.83%	0.5	62.5%	0.25
roll CC	70.83%	0.67	spin left	25%	0.4	50%	0.12
select	87.5%	0.79	glow	20.83%	0.54	48.83%	0.27
shrink	83.33%	0.75	shrink	45.83%	0.25	66.67%	0.25
yaw left	75%	0.79	spin left	33.33%	0.62	37.5%	0.5
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
Gesture+speech							
create	75%	0.33	appear	33.33%	0.18	41.67%	0.15
delete	91.67%	0.33	disappear	54.17%	0.33	37.47%	0.00
enlarge	66.67%	0.29	enlarge	25%	0.56	41.64%	0.27
move away	41.67%	0.44	move back	16.67%	0.64	25%	0.2
move down	58.33%	0.5	drop	29.17%	0.46	29.16%	0.04
move left	70.83%	1	move left	25%	0.2	45.83%	0.8
move right	75%	1	move right	20.83%	0.33	54.17%	0.67
move towards	37.5%	0.33	move forward	16.67%	0.43	20.83%	0.1
move up	66.67%	0.5	move up	41.67%	0.33	25%	0.17
pitch down	79.17%	0.4	spin forward	20.83%	0.6	58.34%	0.2
pitch up	75%	0.17	spin back	16.67%	0.43	58.33%	0.26
roll C	62.5%	0.33	rotate	20.83%	0.36	41.67%	0.03
roll CC	66.67%	0.29	spin left	25%	0.23	41.67%	0.06
select	79.17%	0.67	change	25%	0.36	54.17%	0.31
shrink	75%	0.5	shrink	41.67%	0.23	33.33%	0.27
yaw left	79.17%	0.6	spin	29.17%	0.36	50%	0.24
yaw right	79.17%	0.33	rotate right	20.83%	0.6	58.37%	0.27
Column Average	69.36%	0.47		27.21%	0.39	42.15%	0.24

Legend:*: Top proposal from E-Text is the referent as displayed, C: Clockwise, CC: Counterclockwise, $\mathcal{MC}$: Max-Consensus, $\mathcal{CDR}$: Consensus-Distinct Ratio, differences are absolute values

4.4 Gesture and Speech Initiation Times Comparison

The average times between gesture initiation and speech initiation were normally distributed in E-Text ($W(24) = 0.967, p = 0.603$), but were not in E-Animated ($W(24) = 0.896, p = 0.018$) as supported by Shapiro-Wilk normality tests. Due to some of the data being non-normally distributed, a Wilcoxon Rank Sum test was performed ($W = 469, Z = 6.021, P < .001, r = 67.69$). Which indicated a shift between the median values of the two distributions of times.

In E-Text there was a much longer wait time between when gestures were initiated and when speech was initiated. In E-Animated, the time of speech initiation was around 60 ms faster on average than it was in E-Text further suggesting an impact of referent display on co-occurring gesture+speech proposal generation (Table 4.4).

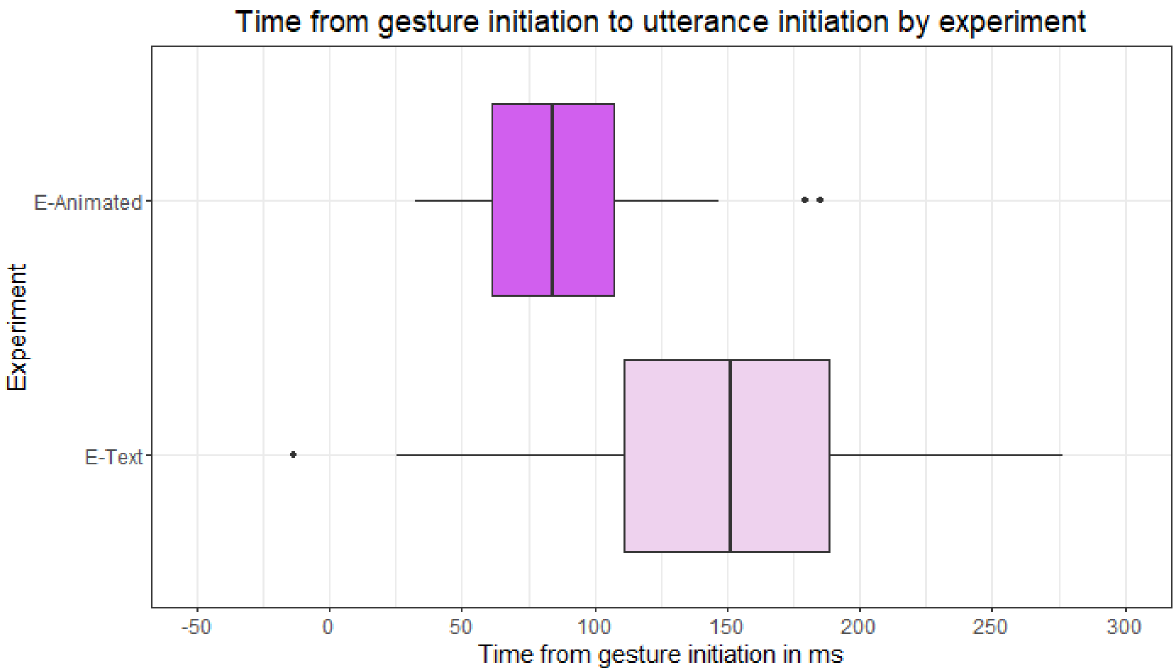

Figure 4.4: Comparison of the time between gesture initiation and utterance initiation in the Gesture+Speech condition

Table 4.4: Time between gesture and utterance initiation by experiment in milliseconds

Experiment	Mean	Median	Standard Deviation
E-Text	151.31	151.19	68.86
E-Animated	90.87	84.32	38.35
Difference	60.44	66.87	30.51

4.5 NASA Task Load Index

The NASA-TLX overall scores for each experiment and input condition are shown in Table 4.5. The NASA TLX results by condition and experiment were normally distributed based on the results of Shapiro-Wilk tests: E-Animated gesture: $W(26) = .932, p = .087$, E-Animated speech: $W(26) = .971, p = .647$, E-Animated gesture+speech: $W(26) = 0.928, p = .07$, 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 ($t(47.926) = 2.633, p = 0.011$, $t(47.379) = 3.18, p = .003$ respectively). This difference was not found for the speech conditions ($t(47.459) = 0.529, p = .6$). These scores can be seen in Figure 4.5.

E-Animated has a lower score than E-Text for each condition (Table 4.5). The gesture+speech condition had the largest difference in perceived workload between studies (13.2) followed by the gesture condition with a difference of 10.6. The difference between the speech conditions is the lowest at 2.3. The difference in perceived difficulty in both the gesture alone or 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 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%) in the speech condition of E-Text.

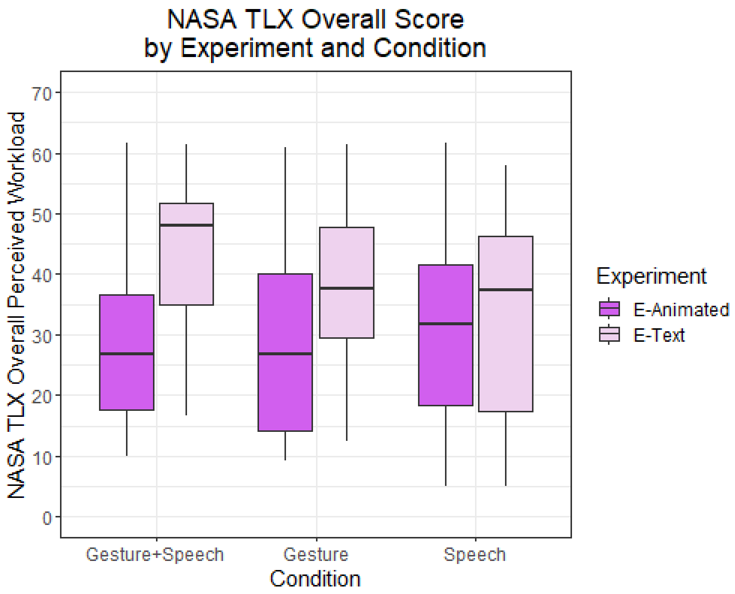

Figure 4.5: Comparison of NASA TLX overall workload by condition and experiment with differences

Table 4.5: NASA TLX overall scores by experiment and condition with absolute differences

	Gesture			Speech			Gesture+Speech		
	E-Text	E-Animated	**Diff**	E-Text	E-Animated	**Diff**	E-Text	E-Animated	**Diff**
Mean	39.3	28.7	10.6	33.5	31.2	2.3	43.5	30.3	13.2
SD	13.4	15.1	1.7	15.6	15.2	0.4	13.3	16.2	2.9

Legend: **Diff**: absolute value difference between experiments

4.6 Consensus Set

Out of the gestures proposed in E-Text, a consensus set of the most highly agreed-upon gesture for each referent was generated. As the $\mathcal{AR}$ decreases, the likelihood that these will be discoverable gestures also decreases. The $\mathcal{AR}$ for these referents is shown in Table 4.6. The gesture proposals from E-Text were chosen for the consensus set due to the limited amount of priming text referents displayed when compared to animated referents.

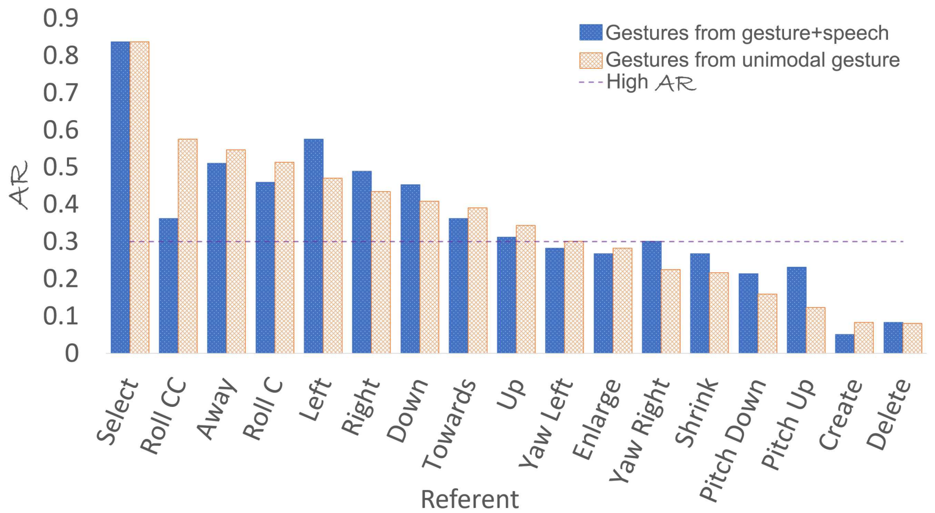

Figure 4.6: Agreement rates for gesture proposals from the gesture block and the gesture proposals from the gesture+speech block for of E-Text (borrowed from [1])

Legend: C: Clockwise, CC: Counter Clockwise

In E-Text, the majority of referents had a clear most agreed-upon gesture proposal (Figure 4.7). These common gestures matched in the gesture condition and the gesture+speech condition. Most of these gestures were symmetric meaning that a "roll right" might be the same gesture as was used for "roll left" with a difference in the direction of the circular motion used. There were some ties where a referent had multiple gesture proposals with equal frequencies (Figure 4.8). An example of these ties is in the *create* referent (Figure 4.8) which had a snap, "bloom", and tap all proposed with equal frequency. Some gestures were more specific to one input modality condition. For the gesture condition participants proposed a finger swipe gesture for delete where in the gesture+speech condition delete was commonly a tap gesture.

A legacy biased gesture called "bloom" that was used on the HoloLens 1 AR-HMD was one of the most proposed gestures for *create*. This may be because 11/24 had prior experience with the HoloLens 1. Another legacy gesture found was the 2-finger zoom-in gesture used on multi-touch phones which was proposed for *enlarge*. *Enlarge* also had a bi-manual expansion gesture tied for the most common proposal.

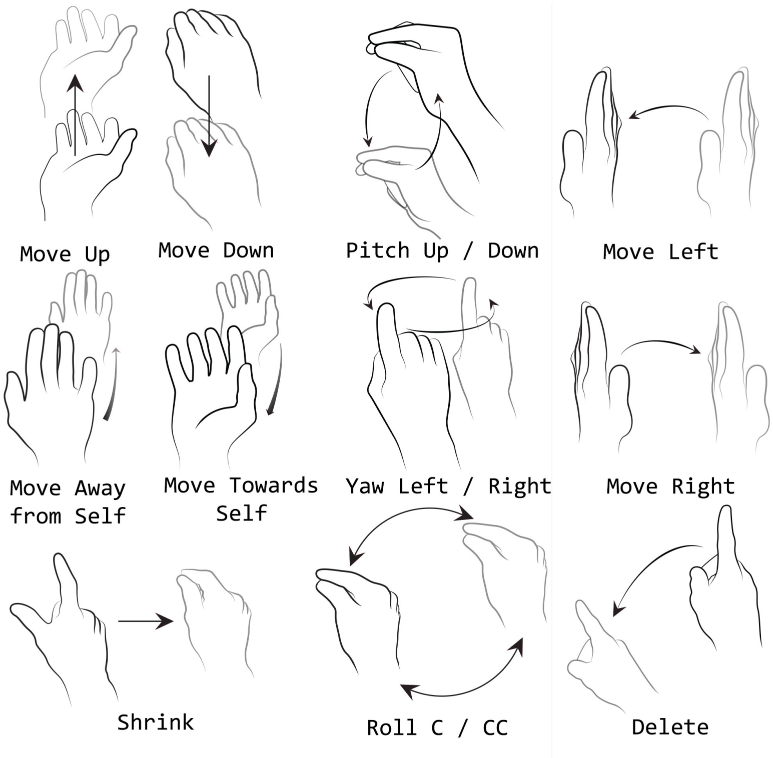

Figure 4.7: Proposed gesture set; C: Clockwise; CC: Counter Clockwise; Bi-directional gestures indicated with double arrows (borrowed from [1])

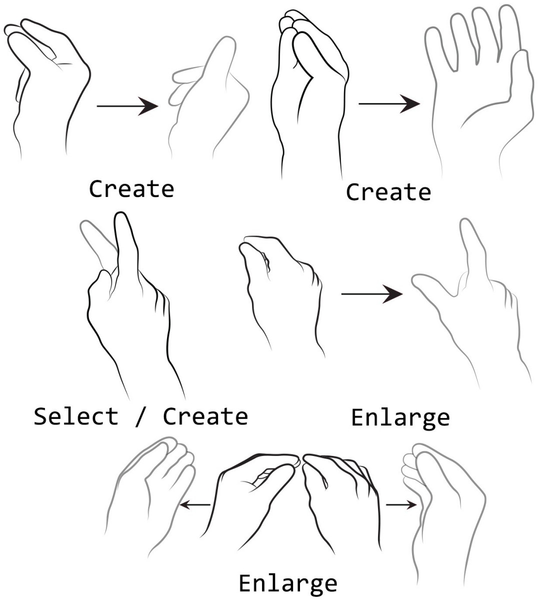

Figure 4.8: Consensus gestures with ties (borrowed from [1])

Chapter 5

Discussion

5.1 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 caused by the referent display for speech elicitation are more salient than what was seen in the gesture proposals. If imitation biased speech proposals are implemented into a system that does not display the same text as the referents that were used these elicited speech commands will be far less discoverable than the study's $\mathcal{MC}$ suggests. 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 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.

The times between gesture initiation and speech initiation show some of the largest disparities in reproduction. There is a large body of work stemming from observation of the timings of co-occurring gesture and speech in human discourse [81, 82]. Within human-computer interaction these times have been observed for basic pointing gestures [74, 83], and interactions in AR [1, 15, 16]. When prompted with text compared to animations, the time between gesturing and speaking increased 166.51%. The NASA TLX results similarly convey that users perceived a higher workload when prompted with text compared to animations in the gesture+speech condition. This is concerning as these windows of time could be implemented into temporal fusion models, and that within the field there is little consistency in the time windows reported [1, 15, 16, 74, 81, 83].

Imitation of speech is considered more difficult than action imitation which helps to explain the lower difference in the NASA TLX scores found for speech [53, 54].

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 4.2). 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 used on touch screen cellular phones. 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 of the top corner up and towards a participant. In E-Animated, this manifested as gesture proposals that had a similar formation as E-Text but used a diagonal movement where E-Text was commonly only in one axis (Figure 4.3).

Previous work exhibits similar indications of interaction bias and imitation. The animations used in prior work are often not directly specified but are rather 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. [7] matches the diagonal motion found in E-Animated [15]. Imitation is also inherent in the foot gestures that presumably mirror the movements of the animation of the avatar in Felberbaum et al.'s work [65], or the direct manipulations for rotation and translations found my Piumsomboon et al. [5].

The effect of imitation is observable in speech elicitation as well. In Morris, 2012, some of the referents received speech proposals that were the referent as it was read aloud [18]. "Open new tab" was proposed for *open new tab*, and "open browser" for *open browser* [18]. 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 when generating proposals, likely making the imitation of speechless likely to inflate the $\mathcal{MC}$ scores through the use of the alternative input channels. The animations could still have primed the gesture proposals, further confounding their results [18, 19].

5.2 Implications for Elicitation Studies

The top gesture proposals were commonly different while the spread of proposals overlapped heavily, evident in the heat-maps for the gesture condition (Figures 4.3, 4.2, 4.1). The bulk of the higher frequency (darker) proposals occurs in both studies exhibited by comparing any pair of columns for the same referent, as with *yaw right* (Figure 4.1). While these top proposals occur in each study, the relative frequencies of their proposals are different. A designer that strictly implements the top-choice interaction for each command would be negatively affected by the biasing caused by referent display if these gestures were proposed based on animations that are not found in the implemented system because the users would form gestures based on the animations that are present in the system. Most elicitation studies recommend aliasing the most common commands [1, 5, 9, 15, 18]. Through aliasing, the overlapping proportions of the proposal space can be captured, offsetting the impact of biasing.

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 where the shape and motion of the gesture remained consistent. This information is likely lost when pre-processing gesture data from the granular bins to the equivalence classes used to compute $\mathcal{AR}$. These limited effects of referent display on elicited gesture proposals is beneficial as prior work that focuses solely on gesture elicitation is minimally impacted by referent display.

The results of the few prior works that perform speech elicitation are more impacted than prior gesture-only elicitation studies [1, 18, 19]. For translations, the top choice (highest $\mathcal{MC}$) interaction was the same for each study potentially due to the simplicity of the referent (i.e., *move* *left*). Rotations present a greater challenge to designers in that the terms used by participants were largely ambiguous, often using the word “spin” to indicate the type of rotation. The abstract referents had the highest levels of imitation bias due to the difficulty found in animating those referents. When animations attempt to capture an abstract action (i.e., *create*, *select*), that animation primes the user’s understanding of the actual task. Examples of this are found with the use of particle effects in E-Animated causing the referent to be understood as the apparition of an object and the use of a hue increase in *Select* being perceived as a wide range of things from “highlight”, or “glow” to “change color.”

Speech elicitation faces two major disadvantages; showing text causes an artificial increase in consensus metrics (Table 4.3), and showing animations causes proposals that deviate from the intended referent (“highlight” for *select*). The gestures, while biased, were more similar between studies. As elicitation continues to be used for novel inputs outside of gesture alone caution should be had that the impacts of referent display are considered when designing those experiments.

5.3 Implications for Multimodal Elicitation

The imitation of referent display is more difficult to resolve in multimodal elicitation. When prompted with text, speech is biased while gestures are not. When prompted with animations, the opposite is true. As elicitation uses continue to expand, 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 providing a suggestion of the granular commands necessary to complete it. Under that approach the steps the user completes could be decomposed to interactions for referents such as *selecting*, *translating*, and *deselecting* objects. This approach removes the bias caused by explicit referent imitation. Similar methods of observing goal completion as opposed to granular action/interaction pairs are more common in information visualization studies [84, 85]. The likelihood of imitation decays over time [86], so studies that delay proposal generation after referent presentation may be able to reduce the impact of imitation bias. Referent-less elicitation is another approach that could remove imitation bias [3].

5.4 Interaction Formation and Imitation

For both referent displays, imitation occurred either via the imitation of action or text. Gestures imitating the motion of an object are easier to produce due to the close coupling of perceptive and proprioceptive channels [47–49]. Evidence of imitation is found in gestures that followed the animation exactly as shown in E-Animated [15] and presumably across other studies [5, 7]. This ease of imitation is further supported by the lower NASA TLX scores in the animation condition (Table 4.5).

Speech imitation was seen as more difficult, supported by the disparity of perceived workload when imitating speech compared to gesture imitation (Figure 4.5). While this imitation still occurs (Table 4.3) [1, 18, 19], it takes longer than action imitation (Figure 4.4) [55] and was considered more difficult (Figure 4.5) [54].

There is strong evidence of imitation in proposal generation. To limit imitation, gestures should be elicited using text. While this would increase the perceived difficulty of generating proposals, it would elicit fewer imitation gestures and offset the double taxation of the visuospatial sketchpad further lowering the task's cognitive load. Speech should be elicited using animation or the proposed goals-based elicitation method when possible, with care given to the abstract referents if an animation is used. The time differences between gesture and speech initiation are lower when producing speech out of animation (Figure 4.4). We believe this is caused by the difficulties of imitating speech [53, 54], and the double taxation of the participant's phonological loop. These issues are offset through animation-based speech elicitation.

With this understanding of processing and execution through imitation, end-users cognitive load can be lessened through intelligent use of information display. If a user is already in an environment with a high amount of visual information, as seen in 3D visualization systems [87], then an auditory prompt would be best. Alternatively, in a task that requires high levels of auditory processing, a visual prompt would be best. Discoverable interactions would include ones that can be directly imitated. These interactions could be speech commands based on the text displayed in menus, or action gestures based on the user's expectations of movements in the system as informed by affordances and prior movements seen.

5.5 Consensus Gestures

Gestures for translation referents had high agreement rates for both the gesture and the gesture+speech conditions. These were often heavily influenced by real-world physics where participants would reach for the object and perform a direct-manipulation. An example of this is reaching out and pushing on the side of the cube to move it to the left. Rotations were either indirect gestures where a participant would trace a circle in the air in the direction of the intended rotation or would grab a corner of the cube and move their hand in a circular path(Figure 4.7). The referents for roll had a high $\mathcal{AR}$ likely because of the clock metaphor found in the name of the referent ("roll clockwise / counterclockwise").

Select had the highest $\mathcal{AR}$. *Select's* heightened $\mathcal{AR}$ was caused by the frequent use of the legacy tap gesture (Figure 4.7). Other legacy biased gestures included the zoom-in/out gesture from multi-touch devices (i.e., cellphones) for the scale referents and the "bloom" gesture from the Microsoft HoloLens 1 for the *create* referent. The presence of this bias can be leveraged to all more transfer knowledge from prior interaction paradigms to AR environments [6, 23, 41].

5.6 Gesture Comparisons with Prior Work

Where this study found a mixture of bi-manual expansions and legacy touchscreen zoom-in/out gestures for the scale referents (Figure 4.8), prior elicitation studies found only bi-manual expansion gestures [7]. Likewise this work found that most of the gestures proposed for were one-handed where bi-manual gestures were more present in other work [7]. This is seen in the translation gesture proposals found in the work done by Khan et al. [7] which were bi-manual direct manipulations and bi-manual path tracing gestures as opposed to the pushing gestures observed here. These differences in results are likely derived from the participant believing they were interacting with a system versus another human or the presentation of the referents to the users. Khan et al. used a 2D screen where this work used a 3D environment [70].

When comparing these results to a mid-air gesture elicitation study that was also done in AR, the translation gestures were similar while rotation and scale gestures were not [5]. The translation gestures from the work done by Piumsomboon et al. were often open-handed pushes as observed in E-Text [5]. For rotation referents, Piumsomboon et al. had most proposals involving a grab and rotate with a participants wrist where this work found a corner grab and rotation with a participants full arm. Proposed scaling gestures also differed. This study most commonly observed legacy zooming gestures and a single bi-manual expansion gesture in response the the referent 'enlarge' as opposed to the bi-manual gestures observed by Piumsomboon et al. [5]. Most gestures in both studies were reversible [5]. Examples of reversible gestures are seen in the rotation and translation gestures in the consensus set provided here (Figure 4.7). These differences could be due to the difference in the referent display. Piumsomboon et al. showed referents as animations of the intended action where this work showed referents as text.

5.7 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. The snapping gestures omission may have been due to the difficulties encountered with animating an object being created or destroyed or the time between that study and the release of the movie. Regardless, its inclusion in E-Text is an interesting example of a gesture that stems from pop culture. We believe that these culturally influenced gestures represent a mechanism for knowledge transfer from other domains into a new environment as is also the case with legacy biased gestures. Societies growing adaptation of speech-enabled assistants (i.e., Alexa, Google) could be a source of other culturally influenced speech base interactions. As an example, “turn on the lights” and “turn on [name of item]” are both common commands within households that use these assistants. When developing a speech interface it may be beneficial to use these pre-learned commands for actions that have a similar function.

5.8 Recognition System Implications

These elicitation studies provide time windows for gesture and speech multimodal fusion systems. The time windows were different for each type of referent display. The window’s accuracy can be increased by taking the range of the two windows or by using the time window that used a referent display that corresponds with the displays in the system where the recognizer is being used.

In Human-computer interaction, a core focus is the improvement of the user experience. These windows allow the development of a user-centric recognition system that prioritizes the user experience. These time windows might not be the most impactful if used with deep neural networks or recognition systems using infrastructures that require large amounts of training data. These windows can instead be used to build a less computationally demanding architecture that can run using an AR-HMD’s limited processing power. When viewing recognition from a user-centric perspective, this lightweight architecture would improve user’s interactions with a system by running in real-time. These time windows limit the delay that a system would have between receiving its first input and either receiving a second input or determining that the single input was a unimodal command. An example would be the system receiving a pointing gesture then waiting to determine if there is a speech command that uses that pointing gesture for context or if the pointing gesture was a unimodal selection command. Limiting the time between the first input and the response of the system would make the user feel like their interactions are more natural and/or that the system is more responsive, either of which would improve the user’s experience.

Chapter 6

Guidelines

6.1 Plan for Realism in Study Context

The more similar the intended use case the elicitation study is, the less the impact of imitation will matter. If eliciting commands for a system that has text icons on menu bars, using the same names for the referents used would allow more transference of the proposed interactions from the elicitation study to the system. We would also advise against using animations when eliciting gestures and text when eliciting speech unless the intended system employs similar animations or text.

6.2 Plan for biasing

Some authors have acknowledged the priming implicit in referents [6], or suggested using an elicitation methodology with no referents [3]. Other work has removed the bias of referents by asking users to self-report their tasks and means of achieving those tasks to inform interaction design [88]. We recommend a guided approach where natural interactions are observed while giving referents as high-level goals that would require the completion of unstated sub-tasks. A goal-based referent for translations could be *sort the objects shown by color*. The video of participants completing that referent would be broken into sub-movements and interaction proposals for analysis. This removes the imitation of explicit referents and animations.

The results of elicitation studies have included valuable insights on human behavior [13,14,41,64] and the interactions found during elicitation have been implemented with positive results [89,90]. The key to generating findings that are usable by designers is to detail the exact methodology used to allow for an understanding of the introduced biases. The context of the experiment should be similar to the intended use of the interactions [3].

6.3 Report Design Choices

It is important to outline the exact methodology used when conducting an elicitation study. Differences in results can emerge from a slight modification to the referent display, gaps in time between studies [39], and exposure to technologies [19]. In addition to the commonly identified design choices such as referents used, count of participants, and previous device exposure, authors need to describe the way referents were displayed, and how long they were shown.

Imitation of referents caused variations in the proposed interactions. When less traditional inputs are elicited (speech, multimodal combinations) those variations in minor aspects of the experiment can lead to far divergent results. Detailed methodology reporting will help designers know under which exact circumstances the proposed interactions will fit and where they may generalize. This context is important to establishing reproducible work within this field [3].

6.4 Report Common Proposals

As mentioned in prior work [1, 5, 9, 15, 18], and as seen here, aliasing is a powerful tool to capture the interactions of a diverse user base. 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 [18, 19], reporting proposals with hand variations included [5, 15], and showing heat-maps of the proposal space (Figures 4.3, 4.2, and 4.1) [15]. Data-sets should be made public or be available on request with appropriate anonymization and ethics committee or internal review board approval.

6.5 Alias Commands

Redundantly mapping interaction techniques to commands (aliasing) is a technique for capturing a larger group of novice user's first choice interactions [1, 9, 15, 18]. The differences in results seen here are focused on the top choice and least common proposals. The middle of the proposal distribution space overlapped. By aliasing, the top N gesture proposals found in elicitation studies will help to counteract the impacts of imitation biased proposals. Several elicitation papers report taxonomies of gesture types [5, 9, 91], or variations in the gestures caused by hand pose [5, 15]. Utilizing those types of observational patterns when aliasing will further increase the adaptability of an interactive system.

Other work has suggested implementing some interactions that seem promising while not being fully tested [88, 92]. This could help reduce the bias by adding more variety to the implemented interactions. The speech syntax between the two studies consistently included most of the key action information, either the <action> or <direction> phrase. Knowing this a word spotting system could be used with common lexically equivalent commands mapped to interactions.

6.6 Leverage Existing Knowledge

Capturing how users form interactions and how that formation relates to the context of the system and the user's mental model of the system is critical to developing a natural feeling input design. A user's understanding of the functionality of the system will likely be informed by the affordances of the system (e.g., a button can be pressed) [93], user understanding of real-world physics (e.g., a cube can be pushed) [94], and legacy interactions (e.g., pinch to zoom). An additional source of prior expectation may be the culturally relevant interactions at the time such as the snap gesture, or speech-based personal assistant commands. These various pieces of interaction formation can be utilized to help establish high levels of transfer knowledge from prior domains lessening how much a user needs to learn to interact with a new system.

Chapter 7

Limitations

The two studies examined in depth here each used a simple set of referents. They also both only showed a single cube. 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. This work is limited by the used referents simplicity and the simplicity of the cube. This work was also limited by the ways referents were displayed as text or animations. These display choices impacted the range of elicited speech in E-Text where participants repeated the referents. The animations caused some speech proposals to be irrelevant to the referent tested as seen with “highlight” being proposed for the referent *select*. Another limitation encountered during this work was the assumptions of animations and referent names that were used by prior work. Efforts were made to ground these assumptions in the design details given in those works. Even so, the assumptions made may not accurately reflect what was used in prior work.

Chapter 8

Conclusions and Future Work

In elicitation, very few works have attempted to replicate findings. All of the pairs of studies examined here failed to directly reproduce previous findings [1, 15, 18, 19]. For the works done within this lab, the study results support that the differences found in interaction proposals were caused by the way that the referents were displayed. Elicitation design is vulnerable to biasing through action and text imitation during input proposal generation. Most elicitation studies have used referent displays that may have encouraged imitation of them, either through animation, or spoken/text prompts. We propose three promising methods to reduce the impacts of imitation: time delay between referent presentation and proposal generation, goal-based referents, or planning for imitation and eliciting through a different channel (i.e., animation for speech).

Most of the differences observed during the comparison of referent designs were found in the elicited speech proposals. These differences are very evident in the heightened $\mathcal{MC}$ scores and the related speech proposals elicited during E-Text when compared to E-Animated. The speech proposal's syntax was less impacted by referent display.

The differences in gesture proposals between experiments were far less pronounced than what was found in the speech proposals. The largest difference was in the abstract referent select, likely caused by the difficulties of animating the concept of selection.

The overlapping proposal spaces can be safely utilized through aliasing inputs and matching the elicitation context to use case context. The changes to elicitation methodology proposed here contribute to the continual improvement of elicitation studies as they transition from being used for unimodal gesture input design to more varied multimodal input design space.

The way people interact with a system is heavily informed by their mental model of how that system works [95]. These models are based on previous experiences with interactions and the affordances of the systems. The affordances of an object are likely to be what is gestured about [96] and what is spoken about [97]. This work contributes to an understanding of how interactions use imitation during referent interpretation and input proposal formation. The time differences in co-occurring gesture and speech interactions and the differences in user-perceived workload caused by the way the referents are presented support this theory of input formation.

When placed in the larger scope of interaction knowledge, a clearer picture of human behavior is painted. Knowing the modality preferences [17–19], the impacts of cognitive load on those preferences [20], and the mechanisms of input formation based on object appearance [96, 97], affordances [95], user mental models [94], and now imitation bias, brings us one step closer to understanding the user. The developer must know their input device and in the case of many natural interaction systems, the input device is the user.

Even so, more analysis of studies where the exact mechanisms of referent display are outlined is necessary to compare to these results. Future work should directly test more of the ways that imitation manifests, comparing across other types of referent display. Some examples of these alternative referent designs are Referent-less design [3, 88], goal-based referents [84, 85], and time delays between referent presentation and input generation. The merits of the goals-based elicitation method and the delayed elicitation method have not been tested. More work is needed to see if they lessen the impacts of referent biasing in elicited proposals.

Work is needed to tell how implementing interactions that were born out of imitation would affect user performance. The proposals found might have been more heavily informed by the affordances of a system than the imitation of aspects of that system. More studies examining animation types and object shapes are needed to untangle the influences of affordances versus animation.

In E-Text a common gesture proposal for *create* was a snapping gesture, presumably caused by the release of the *Marvel Avengers: Endgame* movie at the same time. This snapping gesture represents an interesting cultural artifact. Knowing what other culturally biased interactions are prevalent in society could help to facilitate high levels of transfer knowledge if incorporated into a system. Future work could examine the culturally biased interaction techniques that may arise from the prevalence of interactions with speech-based home assistants (i.e., Amazon Alexa).

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 find if there are ways to form adaptive interfaces that prompt users with specific formats of information to prime the input modality users choose to use to complete that prompt. 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.

Bibliography

[1] Adam S. Williams, Jason Garcia, and Francisco Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization & Computer Graphics*, 2020 In press.

[2] Wilfred J. Hansen. User engineering principles for interactive systems. In *Proceedings of the November 16-18, 1971, Fall Joint Computer Conference*, AFIPS '71 (Fall), page 523–532, New York, NY, USA, 1972. Association for Computing Machinery.

[3] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob 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)*, page NA, Eindhoven, 2020. ACM Press.

[4] Radu-Daniel Vatavu. There's a world outside your tv: Exploring interactions beyond the physical tv screen. EuroITV '13, page 143–152, New York, NY, USA, 2013. Association for Computing Machinery.

[5] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. User-defined gestures for augmented reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 955–960, New York, NY, USA, 2013. Association for Computing Machinery.

[6] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. User elicitation on single-hand microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3403–3414, New York, NY, USA, 2016. Association for Computing Machinery.

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

[8] Jacob O Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A Myers. Maximizing the guessability of symbolic input, 2005.

[9] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1083–1092, New York, NY, USA, 2009. ACM.

[10] James Hollan, Edwin Hutchins, and David Kirsh. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(2):174–196, 2000.

[11] Aurélie Cohé and Martin Hachet. Understanding user gestures for manipulating 3d objects from touchscreen inputs. In *Proceedings of Graphics Interface 2012*, GI '12, pages 157–164, Toronto, Ont., Canada, Canada, 2012. Canadian Information Processing Society.

[12] Sarah Buchanan, Bourke Floyd, Will Holderness, and Joseph 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, page 231–240, New York, NY, USA, 2013. Association for Computing Machinery.

[13] Katherine Tarre, Adam S. Williams, Lukas Borges, Naphtali D. Rishe, Armando B. Barreto, and Francisco 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, pages 125:1–125:2, New York, NY, USA, 2018. ACM.

[14] Tran Pham, Jo Vermeulen, Anthony Tang, and Lindsay MacDonald Vermeulen. Scale impacts elicited gestures for manipulating holograms: Implications for AR gesture design. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 227–240. ACM, June 2018.

[15] Adam S. Williams, Jason Garcia, and Francisco Ortega. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Pre-Print*, 2020.

[16] Minkyung Lee and Mark Billinghurst. A wizard of oz study for an ar multimodal interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, page 249–256, New York, NY, USA, 2008. Association for Computing Machinery.

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

[18] Meredith Ringel 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, pages 95–104, New York, NY, USA, 2012. ACM.

[19] Michael Nebeling, Alexander Huber, David Ott, and Moira 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, page 15–24, New York, NY, USA, 2014. Association for Computing Machinery.

[20] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 129–136, New York, NY, USA, 2004. Association for Computing Machinery.

[21] Mark Micire, Munjal Desai, Amanda Courtemanche, Katherine M. Tsui, and Holly 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, pages 41–48, New York, NY, USA, 2009. ACM.

[22] Jaime Ruiz, Yang Li, and Edward Lank. User-defined motion gestures for mobile interaction, 2011.

[23] 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)*, pages 144–153, Los Angeles, CA, USA, 2017. IEEE.

[24] Ionuţ-Alexandru Zaiţi, Ştefan-Gheorghe Pentiuc, and Radu-Daniel Vatavu. On free-hand TV control: experimental results on user-elicited gestures with leap motion. *Pers. Ubiquit. Comput.*, 19(5):821–838, August 2015.

[25] Thomas Plank, Hans-Christian Jetter, Roman Rädle, Clemens N. Klokmose, Thomas Luger, and Harald 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, pages 4548–4560, New York, NY, USA, 2017. ACM.

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

[27] Aakar Gupta, Thomas Pietrzak, Cleon Yau, Nicolas Roussel, and Ravin Balakrishnan. Summon and select: Rapid interaction with interface controls in mid-air. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, pages 52–61, New York, NY, USA, 2017. ACM.

[28] Bashar Altakrouri, Daniel Burmeister, Dennis Boldt, and Andreas 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, pages 5:1–5:10, New York, NY, USA, 2016. ACM.

[29] Jaime Ruiz and Daniel 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, page 3347–3350, New York, NY, USA, 2015. Association for Computing Machinery.

[30] Radu-Daniel Vatavu and Jacob 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, page 1325–1334, New York, NY, USA, 2015. Association for Computing Machinery.

[31] Radu-Daniel 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, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

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

[33] Francesco Cafaro, Leilah Lyons, and Alissa 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, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[34] Phil Cohen, Colin Swindells, Sharon Oviatt, and Alex 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, page 137–140, New York, NY, USA, 2008. Association for Computing Machinery.

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

[36] Patricia K Kuhl and Andrew 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.

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

[38] 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)*, pages 1723–1728, Osaka, Japan, Japan, 2019. IEEE.

[39] Poorna Talkad Sukumar, Anqing Liu, and Ronald Metoyer. Replicating user-defined gestures for text editing. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, ISS '18, page 97–106, New York, NY, USA, 2018. Association for Computing Machinery.

[40] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, M c Schraefel, and Jacob O Wobbrock. Reducing legacy bias in gesture elicitation studies. *Interactions*, 21(3):40–45, May 2014.

[41] Anne Köpsel and Nikola Bubalo. Benefiting from legacy bias. *interactions*, 22(5):44–47, August 2015.

[42] Adam S. Williams and Francisco R. Ortega. Evolutionary gestures: When a gesture is not quite legacy biased. *Interactions*, 27(5):50–53, September 2020.

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

[44] Emiel Cracco, Lara Bardi, Charlotte Desmet, Oliver Genschow, Davide Rigoni, Lize De Coster, Ina Radkova, Eliane Deschrijver, and Marcel Brass. Automatic imitation: A meta-analysis. *Psychological Bulletin*, 144(5):453, 2018.

[45] Roman Liepelt and Marcel Brass. Top-down modulation of motor priming by belief about animacy. *Experimental psychology*, 2010.

[46] Emma 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.

[47] George Butterworth. On reconceptualising sensori-motor development in dynamic systems terms. In *Sensory-motor organizations and development in infancy and early childhood*, pages 57–73. Springer, 1990.

[48] Jacquelyn T Gray, Ulric Neisser, Beth A Shapiro, and Stephenie Kouns. Observational learning of ballet sequences: The role of kinematic information. *Ecological Psychology*, 3(2):121–134, 1991.

[49] Andrew N Meltzoff and M Keith Moore. Imitation, memory, and the representation of persons. *Infant behavior and development*, 17(1):83–99, 1994.

[50] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental brain research*, 91(1):176–180, 1992.

[51] Luciano Fadiga, Leonardo Fogassi, Giovanni Pavesi, and Giacomo Rizzolatti. Motor facilitation during action observation: a magnetic stimulation study. *Journal of neurophysiology*, 73(6):2608–2611, 1995.

[52] Marco Iacoboni, Roger P Woods, Marcel Brass, Harold Bekkering, John C Mazziotta, and Giacomo Rizzolatti. Cortical mechanisms of human imitation. *science*, 286(5449):2526–2528, 1999.

[53] Andrew J Lotto, Gregory S Hickok, and Lori L Holt. Reflections on mirror neurons and speech perception. *Trends in cognitive sciences*, 13(3):110–114, 2009.

[54] Alvin M Liberman and Ignatius G Mattingly. A specialization for speech perception. *Science*, 243(4890):489–494, 1989.

[55] Maksim Stamenov and Vittorio Gallese. *Mirror neurons and the evolution of brain and language*, volume 42. John Benjamins Publishing, 2002.

[56] Chaklam Silpasuwanchai and Xiangshi Ren. Designing concurrent full-body gestures for intense gameplay. *Int. J. Hum.-Comput. Stud.*, 80(C):1–13, August 2015.

[57] Zhen Chen, Xiaochi Ma, Zeya Peng, Ying Zhou, Mengge Yao, Zheng Ma, Ci Wang, Zaifeng Gao, and Mowei 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.

[58] Gustavo Alberto Rovelo Ruiz, Davy Vanacken, Kris Luyten, Francisco Abad, and Emilio Camahort. Multi-viewer gesture-based interaction for omni-directional video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4077–4086, 2014.

[59] Lina Lee, Yousra Javed, Steven Danilowicz, and Mary Lou Maher. Information at the wave of your hand. In *Proceedings of HCI Korea*, HCIK '15, page 63–70, Seoul, KOR, 2014. Hanbit Media, Inc.

[60] Nem Khan Dim, Chaklam Silpasuwanchai, Sayan Sarcar, and Xiangshi 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, page 204–214, New York, NY, USA, 2016. Association for Computing Machinery.

[61] Lynn Hoff, Eva Hornecker, and Sven 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, page 86–91, New York, NY, USA, 2016. Association for Computing Machinery.

[62] Keenan R. May, Thomas M. Gable, and Bruce 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, page 74–83, New York, NY, USA, 2017. Association for Computing Machinery.

[63] Panayiotis Koutsabasis and Chris K. Domouzis. Mid-air browsing and selection in image collections. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, page 21–27, New York, NY, USA, 2016. Association for Computing Machinery.

[64] Sabrina Connell, Pei-Yi Kuo, Liu Liu, and Anne Marie 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, page 277–280, New York, NY, USA, 2013. Association for Computing Machinery.

[65] Yasmin Felberbaum and Joel Lanir. Better understanding of foot gestures: An elicitation study. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[66] Ionuţ-Alexandru Zaiţi, Ştefan-Gheorghe Pentiuc, and Radu-Daniel 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.

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

[68] 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*, page 363–371, USA, 1988. Cambridge University Press.

[69] Uran Oh and Leah Findlater. The challenges and potential of end-user gesture customization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 1129–1138, New York, NY, USA, 2013. Association for Computing Machinery.

[70] Susan Wagner Cook and Michael K Tanenhaus. Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, 113(1):98–104, 2009.

[71] Gennaro Costagliola, Mattia De Rosa, and Vittorio Fuccella. A technique for improving text editing on touchscreen devices. *Journal of Visual Languages & Computing*, 47:1–8, 2018.

[72] Marie-Luce Bourguet and Akio Ando. Synchronization of speech and hand gestures during multimodal human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, CHI '98, page 241–242, New York, NY, USA, 1998. Association for Computing Machinery.

[73] Sandrine Robbe. An empirical study of speech and gesture interaction: Toward the definition of ergonomic design guidelines. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, CHI '98, page 349–350, New York, NY, USA, 1998. Association for Computing Machinery.

[74] Sylvia Irawati, Scott Green, Mark Billinghurst, Andreas Duenser, and Heedong Ko. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In *Proceedings of the 16th International Conference on Advances in Artificial Reality and Tele-Existence*, ICAT'06, page 272–283, Berlin, Heidelberg, 2006. Springer-Verlag.

[75] Christophe Mignot, Claude Valot, and Noëlle Carbonell. An experimental study of future "natural" multimodal human-computer interaction. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, CHI '93, page 67–68, New York, NY, USA, 1993. Association for Computing Machinery.

[76] Hessam Jahani and Manolya Kavakli. Exploring a user-defined gesture vocabulary for descriptive mid-air interactions. *Cogn. Technol. Work*, 20(1):11–22, February 2018.

[77] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139 – 183. North-Holland, USA, 1988.

[78] Doug A. Bowman, Ernst Kruijff, Joseph J. LaViola, and Ivan Poupyrev. *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., USA, 2004.

[79] Francisco R Ortega, Fatemeh Abyarjoo, Armando Barreto, Naphtali Rishe, and Malek Adjouadi. *Interaction design for 3D user interfaces: The world of modern input devices for research, applications, and game development*. CRC Press, 2016.

[80] Adam S. Williams, Jason Garcia, Fernando De Zayas, Fidel Hernandez, Julia Sharp, and Francisco R. Ortega. The cost of production in elicitation studies and the legacy bias-consensus trade off. *Multimodal Technologies and Interaction*, 4(4), 2020.

[81] David Mcneill. *Gesture and Thought*. the University of Chicago Press, USA, 01 2005.

[82] Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance in m. *The Relationship of Verbal and Nonverbal Communication*, 25, 01 1980.

[83] Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, ICMI '03, page 12–19, New York, NY, USA, 2003. Association for Computing Machinery.

[84] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE, 2005.

[85] Bongshin Lee, Catherine Plaisant, Cynthia Sims Parr, Jean-Daniel Fekete, and Nathalie 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, page 1–5, New York, NY, USA, 2006. Association for Computing Machinery.

[86] Agnes Moors and Jan De Houwer. Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, 132(2):297, 2006.

[87] M. Cordeil, A. Cunningham, B. Bach, C. Hurter, B. H. Thomas, K. Marriott, and T. Dwyer. Iatk: An immersive analytics toolkit. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 200–209, 2019.

[88] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. Pen + touch = new tools. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 27–36, New York, NY, USA, 2010. Association for Computing Machinery.

[89] 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)*, pages 73–82, Munich, Germany, 2014. IEEE.

[90] Yi-Jheng Huang, Takanori Fujiwara, Yun-Xuan Lin, Wen-Chieh Lin, and Kwan-Liu Ma. A gesture system for graph visualization in virtual reality environments. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 41–45. IEEE, 2017.

[91] Celeste Groenewald, Craig Anslow, Junayed Islam, Chris Rooney, Peter Passmore, and William Wong. Understanding 3d mid-air hand gestures with interactive surfaces and displays: A systematic literature review. In *Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion!*, HCI '16, Swindon, GBR, 2016. BCS Learning &; Development Ltd.

[92] Bill Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.

[93] Donald A. Norman. *The Design of Everyday Things*. Basic Books, Inc., New York, NY, USA, 2002.

[94] Robert J.K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. Reality-based interaction: A framework for post-wimp interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 201–210, New York, NY, USA, 2008. Association for Computing Machinery.

[95] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[96] Ingrid Masson-Carro, Martijn Goudbeek, and Emiel Krahmer. How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior*, 41(4):367–394, 2017.

[97] Karen J Pine, Daniel J Gurney, and Ben Fletcher. The semantic specificity hypothesis: When gestures do not depend upon the presence of a listener. *Journal of Nonverbal Behavior*, 34(3):169–178, 2010.

Appendix A

Heat-maps for the Gestures from the Gesture+Speech Condition

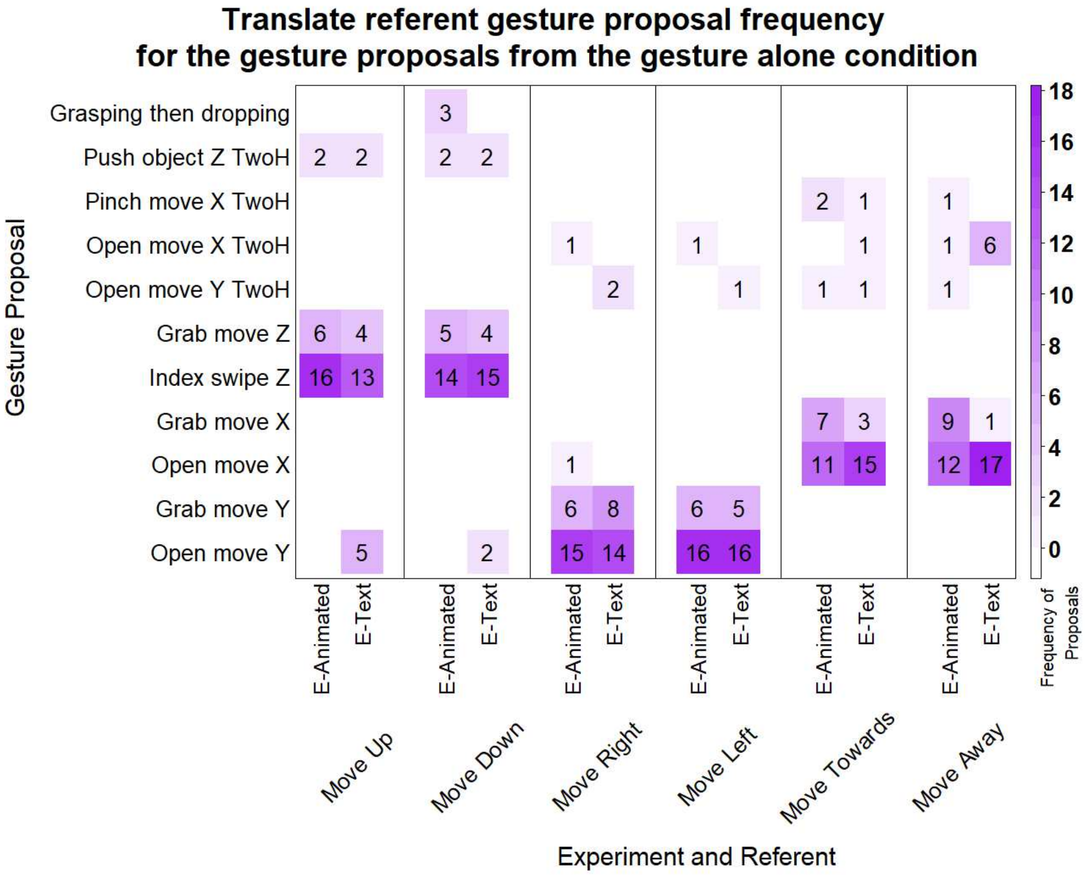

Figure A.1: Heatmap of common gesture proposals by referent and experiment (translation referents only) **Legend**: Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed

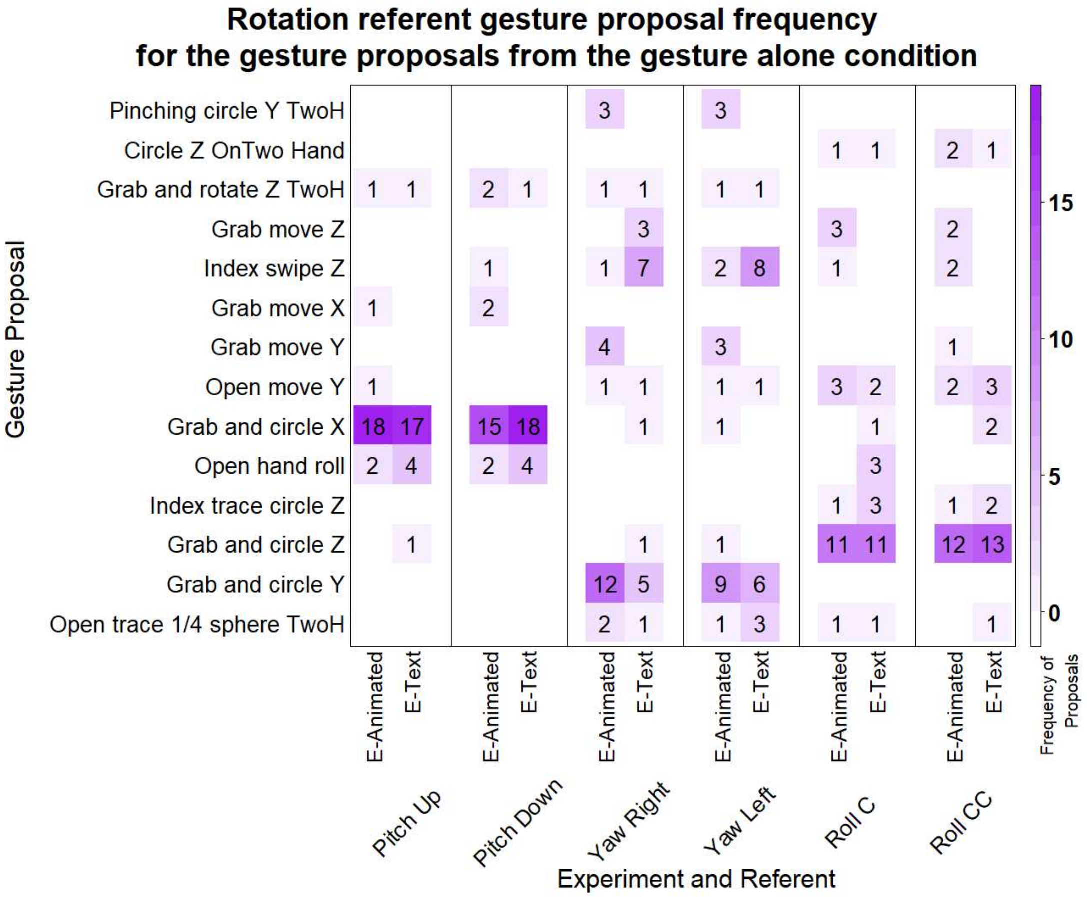

Figure A.2: Heatmap of common gesture proposals by referent and experiment (rotation referents only) **Legend**: E-T: E-Text, E-A: E-Animated, Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed, CC: counterclockwise, C: clockwise

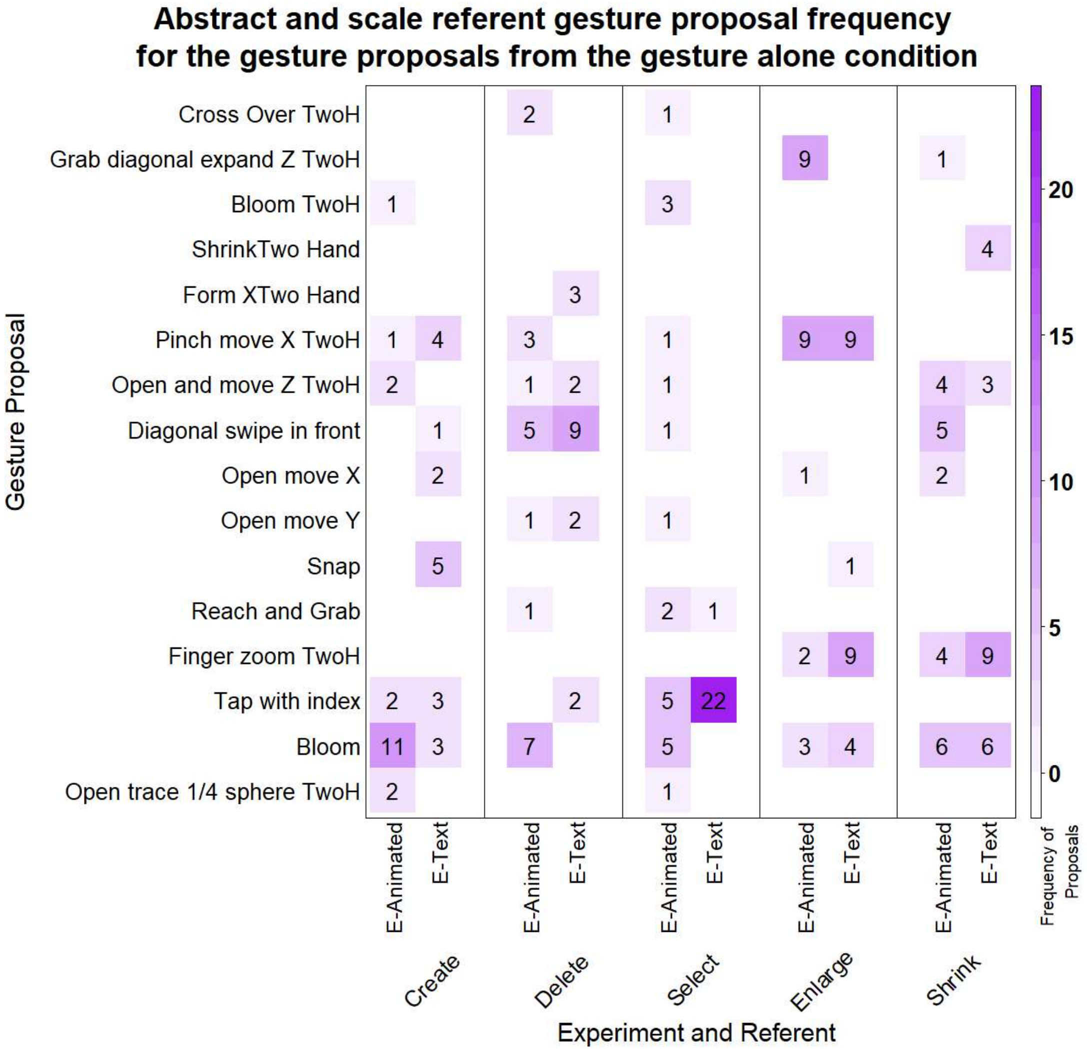

Figure A.3: Heatmap of common gesture proposals by referent and experiment (abstract and scale referents only)

Legend: E-T: E-Text, E-A: E-Animated, Z-axis: vertical, Y-axis: horizontal, X-axis: forward/back, Open: open hand, Grasping: grabbing hand position, Push: open palm push, TwoH: two handed