Selection and Manipulation Whole-Body Gesture Elicitation Study In Virtual Reality

Francisco R. Ortega* Colorado State University
Katherine Tarre† Florida International University
Mathew Kress‡ Florida International University
Adam S. Williams§ Colorado State University
Armando B. Barreto¶ Florida International University
Naphtali D. Rishe‖ Florida International University

ABSTRACT

We present a whole-body gesture elicitation study using Head Mounted Displays, including a legacy bias reduction. The motivation for this study was to understand the type of gesture agreement rates for selection and manipulation interactions and to improve the user experience for whole-body interactions. We looked at 23 participants and 20 distinct referents (with multiple gestures per referent). We found that regardless of the production technique used to remove legacy bias, legacy bias was still found in some of the produced gestures. In some instances, gestures were derived from previous interactions but were still appropriate for the environment presented. This study provides a rich set of information and useful recommendations for future designers and/or developers.

Index Terms: Gesture Elicitation—Gestures—Virtual Reality—Whole-Body;

1 INTRODUCTION

The abundance of Virtual Reality (VR) and head-mounted displays (HMDs) has allowed developers to reach a larger audience. A common interaction technique is the use of VR controllers. While these controllers have grown in popularity (since they come bundled with popular HMDs), there remains the question of how to improve interaction by using Natural User Interfaces (NUIs). The use of gestures provides a more intuitive and expressive form of interaction. However, it is unclear what type of gestures would be ideal for a VR environment. In this paper, we delve into the following questions: (1) What gestures are appropriate for selection and manipulation in a VR environment? (2) Does the size of an object effect how participants interact with it? (3) Would legacy bias still be present after using a reduction legacy bias method? This whole-body gesture elicitation study provides the findings and observations to these questions and additional recommendations.

This user study provides the following **contributions**. First, using the production legacy bias reduction method participants provided different gestures for the environment under consideration that would have not otherwise occur; however, gesture legacy is still apparent in the study. Second, we looked at at the co-agreement rates, which are not always included in gesture elicitation studies. This provides a richer understanding of the results. Third, this study allows for user input (based on their favorite gestures). We believe that understanding selection and manipulation, specifically with different objects is critical for the applications that will be developed in the near future for HMDs and similar devices. Fourth, after the study, we have found some indication that production may not be an effective legacy bias reduction technique. Finally, we have validated that the size of the object may influence the type of gesture, as also found in [12].

*e-mail:fortega@colostate.edu
†e-mail:ktarr007@fiu.edu
‡e-mail:mkres006@fiu.edu
§e-mail:AdamWil@colostate.edu
¶e-mail:Barretoa@fiu.edu
‖e-mail:ndr@acm.org

2 LEGACY BIAS AND ELICITATION STUDIES

Elicitation studies are critical to understand gestures and preferences from users. However, we understand that this practice has generated debate and confusion. For example, legacy bias reduction techniques [9] have been suggested but very little evidence has been shown that these techniques may provide an improved gesture set. In addition, legacy bias in itself can be beneficial [10]. One of the reasons we conducted this study was to understand production and its consequences. There is nothing definitive in the current literature about legacy bias.

Another important aspect about elicitation studies is the methodology used to conduct them. We used the gesture elicitation methodology introduced by Wobbrocks et al. and refined by Vatavu and Wobbrock [17, 18]. We have chosen this methodology because it is currently the most empirically tested gesture elicitation methodology in use. An article was written by Tsandilas that was published after this experiment was completed. The article [16] casts some doubt on the methodology of Wobbrocks et al. [17, 18]. One of Tsandilas main critiques of this methodology is that chance agreement rates may be overly optimistic. However, Tsandilas's paper is a theoretical exercise. He did use some previous studies to validate his arguments. However, the alternatives he offers are not validated [16]. While Tsandilas makes some compelling arguments, we believe that his suggested methodology needs further validation. In particular, it needs more human subjects experimentation done with either the alternative methodologies proposed by him or someone else.

3 RELATED WORK

The main two objectives of a gesture elicitation study are to collect a gesture set from the users and to understand user behavior [18]. The popularity of gesture elicitation is reflected in a variety of studies, ranging from 3D travel using multi-touch and mid-air gestures [10], mobile devices [8], in-vehicle gestures [7], accessibility [3], and multi-touch surfaces [1].

A concern that continues to be critical in gesture elicitation studies is the effect of **legacy bias** on the results. This is not to say that legacy bias is always a problem, as it can be good for transitioning to new devices [5]. Legacy bias originates from the experience that users have with previous technologies (e.g., multi-touch or mouse). Morris et al. suggested steps to reduce legacy bias [9], including production, priming, and grouping [9]. In this study, we have selected to explore production. The production method requires users to produce multiple gestures for a given referent. This approach has been used in different studies. For example, Hoff et al. used it for eliciting mid-air gestures for music playlists [4] and Ruiz et al. used it for for whole-body gestures [14]. Other reduction

2019 IEEE Conference on Virtual Reality and
3D User Interfaces
23-27 March, Osaka, Japan
978-1-7281-1377-7/19/$31.00 ©2019 IEEE

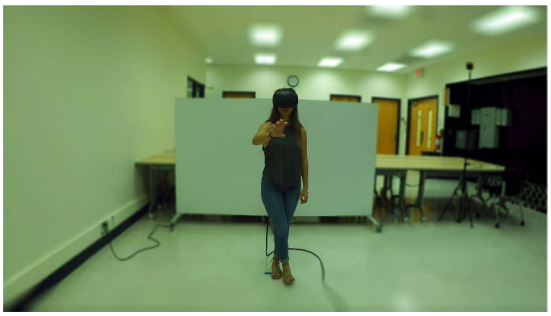

Figure 1: Subject during Gesture Elicitation

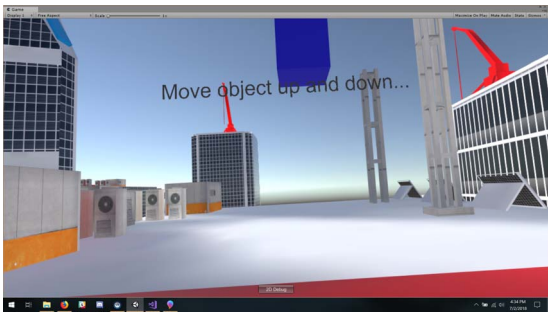

Figure 2: After User has moved object "UP"

methods proposed by [9] were used by Rodriguez and Marquardt to find out how users would opt-in or opt-out from public displays [13].

This is not the first study conducted for whole-body interaction. Connell et al. conducted a whole-body study with children [2]. Besides the common problem found with legacy device bias, they found low-agreement rates, which provides some similarities to our study. A different study [15] for whole-body intense play, used a different strategy for one of their experiments: choice-based elicitation, proposed by the authors, wherein they provided a predefined-list of gestures (allowing for some creativity). In that design, even when the users selected gestures from a predefined list, the agreement rates were still low (and lower overall, including their gesture elicitation) [15]. Additional whole-body studies include [11].

A study by Koutsabasis and Domouzis examined gesture elicitation for mid-Air interactions [6]. This study used a desktop environment to browse and select images. While the experiment's main objective was to search for mid-air interactions, some of the gestures utilized other parts of the body, making it similar to whole-body interaction [6]. Some of the commonalities between this study and ours are the types of gestures recorded, such as swipe and push (in this and the present study, the likelihood of these types of gestures is higher than others) [6]. This study also utilized two legacy bias reduction methods (production and priming), the effectiveness of which is not clear [6]. This remains an open question in our view (see Discussion Section).

Pham et al. looked at eliciting gestures in an augmented reality environment (wearing a Microsoft Hololens) [11]. The major commonality with this study and ours is the overlap of gestures across multiple referents.

4 User Study and Evaluation

When designing the experiment, several decisions were made. First, based on previous literature, the production method for legacy bias reduction was selected. Another decision was to suggest three gestures per referent while allowing the participant to choose more or less gestures, with one as the minimum requirement. This was done to allow more expressiveness in the gesture discovery process. We chose to use **direct manipulation** for objects in this environment. In a actual recognition system, transfer functions would have to be implemented. A Microsoft Kinect version 2 was added to record additional data. We also used the Kinect as a prop by telling participants that it was recognizing their gestures.

We used Windows 10, Unity Game Engine 2017.1, HTC Vive HMD (without the controllers), and Microsoft Kinect version 2. The Kinect was used to record data but not to recognize gestures. We also use GOPro 4 camera to record the user gestures for analysis. The experimenter made note of the gestures in addition to the recorded data already mentioned.

Each participant was given the IRB inform consent before starting the experiment. Once informed consent was completed, the experimenter asked the subject to complete an entry questionnaire. This questionnaire was design to find basic characteristics such as age, gender, ethnicity, and device experiance (e.g., have you used the Microsoft Kinect). In the initial phase, participants were given 5 minutes to get used the environment while wearing the HMD. Then, subjects were asked to performed an elicitation for two referents as part of their training. These were different from the actual referents used during the experiment. Once that was concluded, referents were asked in randomized order. For each referent subjects were asked to provide multiple gestures (production). Each time they performed the gesture, the environment would execute the movement (e.g., an object was created). Once they completed the production phase for a given referent, the experiment would ask them to rate their preference among the gestures produced. This continued until all referents were seen, providing some space for rest when needed. The one problem that was encountered during production was that subjects would fail to produce the three required gestures as asked by the experimenter. In that case, the experimenter would move to the next referent. At the end of the experiment an exit-questionnaire was conducted.

The 23 participants considered were composed of 10 women and 13 men with an average age of 21 years. All of our participants were right-handed and just over 50% of them had no experience with either Microsoft Kinect or VR HMDs. When asked to rate the clarity of the instructions on a scale of 1 to 7, over 80% of the participants said they were very clear with a score of 7.

Participants were asked a series of questions on a scale of 1 to 7, with 1 being the minimum and 7 being the maximum. When asked how accurately the Microsoft Kinect recognized their movements in virtual reality, over 70% of participants responded with scores of 6 or 7. When asked if they enjoyed using the Vive headset, over 80% of the participants scored 6 or 7.

For the exit questionnaire, participants were once again asked a series of questions on a scale of 1 to 7, with 1 being the minimum and 7 being the maximum. These questions dealt with the clarity of the instructions provided (with all participants rating 6 or 7), accuracy of the Kinect's recording of the VR movement (over 90% of participants rated 5 or above), participants' level of enjoyment in using the Vive headset (Over 80% of participants rated 6 or above), extent of participants' mental immersion experience (95% of participants rated 3 to 5), general enjoyableness of the experiment (over 95% rated 5 or above), level of the environmental responsiveness to the initiated actions (70% of participants rated 6 or above), gesture performance success (all participants rated 4 or above), and gesture performance feeling (80% of participants rated 4 or 5).

4.1 Dataset

The dataset consists of over 1,000 gestures obtained from a total of 30 participants. Participants were encouraged to provide multiple gestures and then select their favorite for each referent. Referents, such as rotate, move, destroy, and create were seen twice using visualization of a cube and a wall, in order to see how participants

would respond to different types of items in the same environment. Our objective was to see if the users would be consistent with their gestures when interacting with items of different sizes, given that the cube was significantly smaller than the wall. In some instances, users were unable or unwilling to develop a gesture for the referent. For this reason, a reduced dataset was used for analysis, including only 23 participants (out of 30) who provided at least one gesture for all referents.

This reduced dataset considered gestures identified as favorites by the participants for each referent. We derived **Overall** agreement, which considered the most repeated gestures (as if it was a gesture elicitation study without production). Finally, we considered the most repeated gestures using the entire set of choices by the users (production).

Overall Agreement rate was derived using a gesture set that only considers one gesture per participant. However, rather than using what the participants select as their favorites, we give preference to the 2 most repeated gestures, followed by their favorite. For example, if a participant provides 3 gestures for a referent, then the *gesture with the highest count* is selected. If none of the provided gestures are in the set of the most repeated gestures, then the choice falls back to what the participant selected as their favorite gesture. The reason we chose to default to the favorite gesture (for this reduced set) is that as the remaining sample of gestures decreases, a larger number of unique gestures is observed, and that allows for the consistent derivation of this table. It is important to note that the three data sets provided are derived from one study.

The gesture agreement was considered using the new formula by Vatavu and Wobbrocks [17] because in the former approach [18], the formula held that even with zero agreement, gestures agreed with themselves. In other words, referents with zero agreement did not have an agreement rate of zero. While this does not invalidate studies that have used the previous approach [18], it is an important factor to consider for comparison. The new approach will have lower agreement rates. In addition, the former approach does not account for the effects of sample size. The analysis for agreement rate was calculated using Formula 1 as proposed by [17].

$$AR(r) = \frac{|P|}{|P|-1} \sum_{Pi \subseteq P} \left(\frac{Pi}{P}\right)^2 - \frac{1}{|P|-1} \tag{1}$$

This study revealed a surprising amount of variation in the agreement rates for the different referents. Table 1 considers the full gesture sets for each referent. The agreement rates range from 4.43%, for CreateCube, to 29.15% , for Z Scale. Interestingly, in most cases, the agreement rate for the representation of the cube seemed to be higher than that of the wall. A possible explanation of this is that due to the cube's dimensions, it was easier for participants to maneuver it in the virtual environment, whereas the wall required more creative management given the implications of its size. When we group the referents (shown in gray rows in Table 1), without consideration for visual effects used, it is clear that agreement rates for X, Y, and Z Move referents are higher than those for rotations. This has manifested in other gesture elicitation studies, such as [10]. In addition, translation (e.g., using swipe gestures) gestures are common in existing devices (e.g., iPhone) and are an indication that production may not always remove legacy bias (see Discussion section). Table 1 provides the amount of collected gestures for each referent (column 2).

Note that the production techniques to mitigate legacy bias create a sense of disagreement within our participants. That is to say that by encouraging each participant to develop multiple gestures for each referent, the disagreement rates increase within the participant itself. Hence, two reduced gesture sets were also analyzed: first, using the gestures participants identified as their favorite of those they developed, and second, by giving preference to the most repeated gestures (called **Overall** agreement).

Table 2 shows the results of the gesture set which considers participants' favorite gestures. In this set, all referents have 23 gestures, one per participant, with the exception of the combined referents, which have 46 observations (each row made of a group in Table 2 is highlighted in gray). It can be observed that the agreement rates have increased significantly with the highest agreement rate for Z Scale rising to 67.98%. Previously discussed was a pattern of cube referents having higher agreement rates than wall referents. However, this no longer holds true in the reduced gesture set.

Table 3 shows the results of the gesture set which considers the most repeated gestures (**Overall** agreement). In this set, all referents have 23 gestures, one per participant, with the exception of the combined referents, which have 46 observations. The gestures for each participant were grouped, and preference was given to the gestures that were repeated the most. In the case that a participant did not propose either of the two most repeated gestures, we defaulted to the one they selected as their favorite. This was not the case in most instances. In this particular reduced dataset, the agreement rates increased significantly with the highest agreement rate for Z Scale, rising to 75.5%. Once again, the previously identified pattern of cube referents having higher agreement rates than wall referents does not hold true in this reduced gesture set. An argument could be made that, because the wall appears more difficult to maneuver, participants fall back on their legacy gestures, which accounts for why so many legacy gestures were seen. However, in the production set, participants were pushed to think outside the box, thereby creating more disagreement in the wall referents.

$$CR(r_1, r_2) = \frac{\sum_{i-1}^{n} \delta_{1,i} \cdot \delta_{i,2}}{n}, Where\ n = \frac{1}{2}|P|(|P|-1) \tag{2}$$

4.1.1 Co-Agreement Rates

To further explore the implications of the relationship between referents using the cube and the wall, simplified co-agreement rates proposed by Vatavu and Wobbrock [17] were utilized, as shown in Formula 2.

Table 4 references the agreement rate for each referent and the resulting co-agreement between cube and wall visualizations using Equation 2 for the "favorites" reduced gesture set. Note that the co-agreement rates for X and Y Move referents are equivalent even though X Move agreement rates are slightly lower than those for Y Move. Z Move has the highest co-agreement rate of 11.07%, even though it does not have the highest agreement rates.

The co-agreement rates for the reduced gesture set, based on most repeated gestures, are shown in Table 6. Note that the co-agreement rates are higher for this gesture set. While the Move referents in the X, Y, and Z planes still have the highest co-agreements, in this set the Destroy referent also has a significantly higher co-agreement rate of 31.23%.

4.1.2 Testing Significance

While the current analysis has given an overview of agreement rates for the presented referents, this does not describe the level of significance in the co-agreement results. Using Vatavu and Wobbrock's extension of Cochran's Test [17], as shown in Formula 3 ($where\ n = \frac{1}{2}|P|(|P|-1)$), we can test whether the hypothesis that the agreement rates of multiple referents remains the same.

$$V_{rd} = n \cdot \frac{(AR(r_1) - AR(r_2))^2}{AR(r_1) + AR(r_2) - 2 \cdot CR(r_1, r_2)} \tag{3}$$

The test statistic (V_{rd}[1]) is then compared to the chi-squared quartiles, with degrees of freedom equal to the number of referents being

[1] "Notation V in V_{rd} stands for the variation between agreement rates, and the subscript rd denote a repeated measures design." [17]

tested minus one, for the significance level being tested. For the purpose of this analysis, the level of significance used was 95%. Given that only pairs of referents were considered, the critical value for the test is 3.84. This means that observations greater than the critical value (3.84) reject the null hypothesis, and we can conclude that the agreement rates for the paired referents are significantly different.

The results of the analysis for the favorite gestures are presented in Table 5. These results show significant differences in the agreement rates of rotations on the Y axis, moves on the Y axis, and moves on the X axis. Furthermore, this analysis can be extended for the same testing procedure to a single referent, and this will allow us to test whether the agreement rate for said referent is significantly different than zero. The results are displayed in Table 4. All individual agreement rates were greater than the critical value; therefore, we can conclude that the individual agreement rates are significantly higher than zero at a 95% level of significance.

The same analysis is conducted for the **Overall** reduced gesture set (considering most repeated gestures). Results are shown in Table 7. Agreement rates for all individual referents are also significantly higher than zero, at a 95% confidence level. Similarly to the favorites gesture set, significant differences between cube and wall referents were found in some of the same cube and wall referents; however, significant differences in the Destroy and Z Move referents were found for this test, as well.

Table 1: Production Agreement

Referent	# G	Gesture	Count	A.Rate
Z_Scale	39	Accordion	20	29.15%
Z_RotateWall	58	2H-Steering Wheel	14	12.28%
Z_RotateCube	53	Doorknob	13	15.09%
Z_Rotate	108	2H-Steering Wheel	26	14.45%
Z_MoveWall	52	2H-Push	20	20.21%
Z_MoveCube	54	2H-Push	16	18.03%
Z_Move	96	2H-Push	36	23.55%
Y_Scale	55	2H-Vertical Accordion	16	16.43%
Y_RotateWall	52	Swipe Right	11	11.61%
Y_RotateCube	56	Swipe Right	19	15.45%
Y_Rotate	101	Swipe Right	30	15.13%
Y_MoveWall	53	2H-Swipe Down	17	18.14%
Y_MoveCube	46	Swipe Down	18	23.96%
Y_Move	98	Swipe Down	32	21.42%
X_Scale	49	Push	12	14.20%
X_RotateCube	54	Swipe Down	11	8.04%
X_RotateWall	58	Push	10	8.41%
X_Rotate	109	Swipe Down	19	8.38%
X_MoveCube	50	Swipe Right	17	14.20%
X_MoveWall	52	Swipe Right	20	17.57%
X_Move	101	Swipe Right	37	16.87%
Select	61	Push	22	16.23%
Destroy_Cube	83	Kick	16	8.90%
Destroy_Wall	74	Punch	15	10.18%
Destroy	155	Kick	31	9.59%
Create_Wall	58	Push	7	4.72%
Create_Cube	53	Push	6	4.43%
Create	104	Push	13	5.77%

Legend: # G: Number of Gestures; A. Rate: Agreement Rate.
Light gray rows are grouping of referents.

Table 2: Favorite Agreement

Referent	# G.	Gesture	Count	A. Rate
Z_Scale	23	Accordion	19	67.98%
Z_RotateWall	23	2H-Steering Wheel	8	18.58%
Z_RotateCube	23	Doorknob	8	20.16%
Z_Rotate	46	Steering Wheel	13	18.60%
Z_MoveWall	23	2H-Push	12	30.04%
Z_MoveCube	23	Push	10	29.25%
Z_Move	46	2H-Push	20	30.00%
Y_Scale	23	2H-Vert. Accordion	11	28.46%
Y_RotateWall	23	Swipe Right	5	11.86%
Y_RotateCube	23	Swipe Right	11	27.27%
Y_Rotate	46	Swipe Right	16	17.20%
Y_MoveWall	23	Swipe Down	9	24.11%
Y_MoveCube	23	2H-Swipe Down	11	39.53%
Y_Move	46	Swipe Down	19	31.98%
X_Scale	23	Push	8	22.13%
X_RotateCube	23	Swipe Down	6	9.09%
X_RotateWall	23	2H-Swipe Down	5	7.91%
X_Rotate	46	Swipe Down	9	8.99%
X_MoveCube	23	Swipe Right	12	32.41%
X_MoveWall	23	Swipe Right	10	22.53%
X_Move	46	Swipe Right	22	28.31%
Select	23	Push	17	54.94%
Destroy_Cube	23	Push	3	5.14%
Destroy_Wall	23	Punch	4	6.72%
Destroy	46	Punch	7	6.28%
Create_Wall	23	Draw Square	4	5.14%
Create_Cube	23	Push	5	8.70%
Create	46	Push	7	7.42%

Legend: # G: Number of Gestures; A. Rate: Agreement Rate.
Light gray rows are grouping of referents.

5 Discussion

One of the most noticeable facts in Tables 1, 2, and 3 is the presence of gestures that would be considered legacy, such as Swipe. However, a few gestures not always found in typical devices (e.g., iPhone or iPad) were present. One of them is the accordion gesture. While this gesture may be considered an evolution from the legacy pinch gesture, the movement was very appropriate for the environment presented in this study. The steering wheel also presents a similarity with the rotate gesture found in tablet devices. However, the movement and motion differs and is also considered appropriate for vision-based recognition of skeletons using a device like the Microsoft Kinect. This is the same case for the Doorknob gesture, as it represents a rotation. Other interesting gestures where Push (ideal for selection) and Punch (ideal for destroy). Nevertheless, it is not entirely clear that production reduces legacy bias when we look at the overall pattern of gestures derived from this study. Another study found that production did not reduce legacy bias [6]. In the case of both this study and [6], there is not enough evidence to say definitively, but there is a trend towards finding that production is not an ideal legacy bias reduction technique. The conclusion is that legacy bias may have been reduced with production in our experiment only in a few gesture but overall legacy seems to be present still. Nevertheless, legacy bias can be useful.

Another observation is the variation between agreement rates with a plurality of them having low agreement rates. While Z Scale had the highest agreement rates for the Favorite and Overall gesture set, this wasn't the case for the X and Y axes. A possible explanation is that it is harder for users to think in 3D, in particular in a true 3D stereoscopic system, such as those rendered in HMDs. For Production agreement rates, the rates involved with cube referents were higher than those dealing with the walls (except for X Move). This could be due to that the cube was significantly smaller than the wall and it was cognitively easier for participants to interact with.

When looking at the co-agreement rates, Move has the highest. This is likely the result of legacy bias, given that participants are most familiar with move referents. Destroy has a high co-agreement rate in the Overall gesture set. This means that people who agreed on a certain gesture for Destroy Cube also agreed on a specific gesture for Destroy Wall, although it is not necessarily the same one.

When looking at the whole-body interaction, while the users were given the choice to use any part of the body and move around (in a small area), in most cases they preferred mid-air gestures. Including clapping hands, double hammer-fist, shoulder bump, shoulder shrug, elbow bump, bow and arrow, among others. Only a few gestures

made use of other body parts. Some of them included knee rise, spread legs, jump and lift, turn head, and head bump, among a few others. The only non-mid air gesture that made it to any of the gesture sets presented here was the the Kick gesture in Table 1. It may feel more natural to use the hands versus other parts of the body.

There are limitations to this study. Even after performing a legacy bias reduction method, legacy bias is still present in some instances. Another limitation, by design, is that the gestures produced here may not translate into a sitting position, where users will be constrained to certain movements.

Table 3: Agreement of Gestures (Overall)

Referent	# G.	Gesture	Count	A. Rate
Z_Scale	23	Accordion	20	75.49%
Z_RotateWall	23	2H-Steering Wheel	14	39.13%
Z_RotateCube	23	Doorknob	12	40.32%
Z_Rotate	46	2H-Steering Wheel	14	22.90%
Z_MoveWall	23	2H-Push	20	75.49%
Z_MoveCube	23	2H-Push	16	53.36%
Z_Move	46	2H-Push	36	63.57%
Y_Scale	23	2H-Vert. Accordion	16	53.36%
Y_RotateWall	23	Swipe Right	11	28.46%
Y_RotateCube	23	Swipe Right	19	67.98%
Y_Rotate	46	Swipe Right	30	43.77%
Y_MoveWall	23	2H-Swipe Down	16	53.36%
Y_MoveCube	23	Swipe Down	18	62.85%
Y_Move	46	Swipe Down	24	45.12%
X_Scale	23	Push	11	33.20%
X_RotateCube	23	Swipe Down	11	30.04%
X_RotateWall	23	Push	9	23.32%
X_Rotate	46	Swipe Down	18	22.13%
X_MoveCube	23	Swipe Right	17	54.94%
X_MoveWall	23	Swipe Right	20	75.49%
X_Move	46	Swipe Right	37	65.22%
Select	23	Push	21	83,00%
Destroy_Cube	23	Kick	16	53.36%
Destroy_Wall	23	Punch	15	42.29%
Destroy	46	Kick	18	27.15%
Create_Wall	23	Push	7	14.62%
Create_Cube	23	Push	6	12.25%
Create	44	Push	13	14.27%

Legend: # G: Number of Gestures; A. Rate: Agreement Rate.
Light gray rows are grouping of referents.

Table 4: Co-Agreements of Favorite Gestures

Referent	Cube A.	Wall A.	Co-A.
Create	8.70%	5.14%	1.19%
Destroy	5.14%	6.72%	0.79%
X_Rotate	9.09%	7.91%	1.19%
Y_Rotate	27.27%	11.86%	3.16%
Z_Rotate	20.16%	18.58%	2.77%
X_Move	32.41%	22.53%	7.91%
Y_Move	39.53%	24.11%	7.91%
Z_Move	29.25%	30.04%	11.07%

Legend: A.: Agreement; Co-A.: Co-Agreement

5.1 Unique Gesture Set?

While it is possible to merge data found in Tables 1, 2, and 3 with the additional data we have (gestures that came in 2nd or 3rd place), it would be an artificial gesture set. Our decision was against it. Nevertheless, we can see some common gestures emerged. For example, for Scaling the Accordion gesture was a favorite with a high agreement rate. The steering wheel and doorknob gestures were common for Z Rotate. A complete summary can be derived by looking at Tables 1, 2, 3.

Table 5: V_{rd} Test for Agreement for Favorite Gestures

Referent	Cube Vrd	Wall Vrd	CW Vrd
Create	22	13	2.793
Destroy	13	17	0.615
X_Rotate	23	20	0.243
Y_Rotate	69	30	**18.325**
Z_Rotate	51	47	0.190
X_Move	82	57	**6.313**
Y_Move	100	61	**12.570**
Z_Move	74	76	0.043

Legend: Table value for 1 DOF: 3.84
Gray rows and **Bold** cells means significant

Table 6: Co-Agreements of Overall Gestures

Referent	Cube A.	Wall A.	Co-A.
Create	12.3%	14.6%	3.56%
Destroy	53.4%	42.3%	31.23%
X_Rotate	30.0%	23.3%	5.93%
Y_Rotate	68.0%	28.5%	20.55%
Z_Rotate	40.3%	39.1%	20.16%
X_Move	54.9%	75.5%	42.69%
Y_Move	62.8%	53.4%	32.41%
Z_Move	53.4%	75.5%	49.80%

Legend: A.: Agreement; Co-A.: Co-Agreement

Table 7: V_{rd} Test for Agreement for Overall Gestures

Referent	Cube Vrd	Wall Vrd	CW Vrd
Create	31	37	0.72
Destroy	135	107	**9.333333**
X_Rotate	76	59	2.752381
Y_Rotate	172	72	**71.42857**
Z_Rotate	102	99	0.090909
X_Move	139	191	**23.7193**
Y_Move	159	135	**4.430769**
Z_Move	135	191	**42.37838**

Legend: Table value for 1 DOF: 3.84
Gray rows and **Bold** cells means significant

5.2 Lessons Learned: Do We Need Production?

Legacy bias has been a point of contention in gesture elicitation. As previously discussed, one of the techniques recommended to reduce legacy bias is to use production. The question still remains if we should even care about legacy bias? When we started the study, we were firmly convinced that some reduction technique was needed. After the study was conducted and with the data presented, we are not so sure if worrying about legacy bias is important. The question of gesture elicitation in itself still requires more study. This goes beyond gesture elicitation but elicitation of behavior for 3D user interfaces.

We have shown that producing a unique gesture set is not always feasible when you have complex environments. It is even more difficult to force subjects to product a number of gestures without them becoming disengaged. Nevertheless, production did create some interesting gestures. For example, the use of the wheel metaphor, the accordion, or the wave back gestures.

6 Conclusion and Future Work

After all the results and discussion, one question remains. What does this mean for user interface designers and/or developers? When designing virtual environments, it is suggested that the Push gesture is used for selection. For moving, swipe is a very standard form of interaction. For rotation, the steering wheel provides a great metaphor but the doorknob is a simpler gesture. For creation and destruction, the drawing metaphor and Punch gesture (respectively)

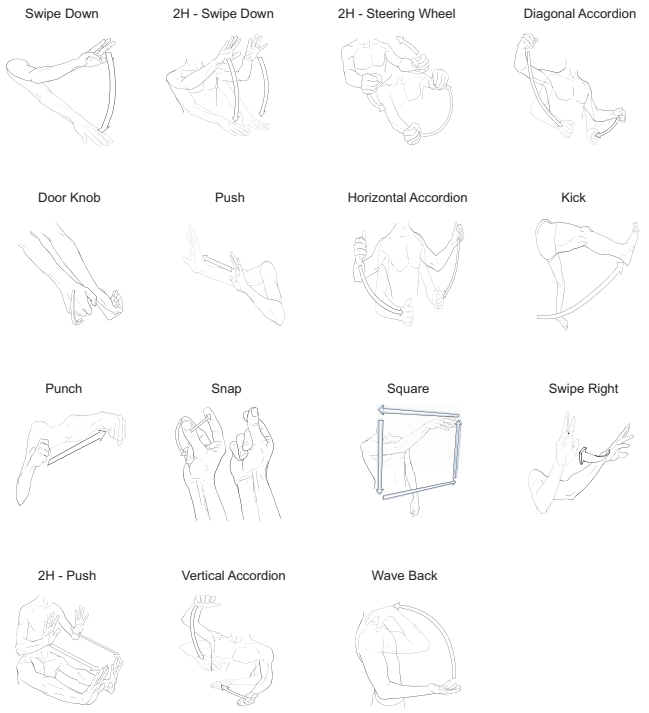

Figure 3: Visual Representation of Most Common or Interesting Gestures of Study

seem to be great options. This may not apply to all domains.

This study delved into gesture elicitation for Virtual Reality for selection and manipulation interaction techniques. The study yielded suggestions for designers and/or developers. The gesture set derived from this study could be implemented in any virtual environment where direct manipulations of objects with varying sizes are needed. In addition, it was found that even after using production, some legacy bias was still present in some of the gestures found in the presented set. In other words, the study yielded some gestures with similarities to multi-touch gestures (but not identical). The accordion and steering wheel gestures present some similarities but were appropriate for the environment presented. The push and punch were different. We found some indications that people will generate different gestures for the same referent depending on the size of the object. This effect has also been found by Pham et. al [12].

There are multiple ways to move this study forward. In particular, the question about legacy bias needs to be explored further. What other methods are there to reduce legacy bias? When do we need to reduce legacy bias? However, the question remains for future work: Should legacy bias be analyze and is the current accepted methodology optimal?

ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Nos. I/UCRC IIP-1338922, III-Large IIS-1213026, MRI CNS-1429345, MRI CNS-1532061.

REFERENCES

[1] S. Buchanan, B. Floyd, W. Holderness, and J. J. LaViola. Towards user-defined multi-touch gestures for 3D objects. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*. ACM, 2013.

[2] S. Connell, P.-Y. Kuo, L. Liu, and A. M. Piper. A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface. In *Proceedings of the 12th International Conference on Interaction Design and Children*, IDC '13, pp. 277–280. ACM, New York, NY, USA, 2013.

[3] N. K. Dim, C. Silpasuwanchai, S. Sarcar, and X. Ren. Designing Mid-Air TV Gestures for Blind People Using User- and Choice-Based Elicitation Approaches. In *the 2016 ACM Conference*, pp. 204–214. ACM Press, New York, New York, USA, 2016.

[4] L. Hoff, E. Hornecker, and S. Bertel. Modifying Gesture Elicitation: Do Kinaesthetic Priming and Increased Production Reduce Legacy Bias? In *TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2016.

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

[6] P. Koutsabasis and C. K. Domouzis. Mid-Air Browsing and Selection in Image Collections. In *the International Working Conference*, pp. 21–27. ACM Press, New York, New York, USA, 2016.

[7] K. R. May, T. M. Gable, and B. N. Walker. Designing an In-Vehicle Air Gesture Set Using Elicitation Methods. In *the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 74–83. ACM Press, New York, New York, USA, 2017.

[8] S. N. Medrano, M. Pfeiffer, and C. Kray. Enabling remote deictic communication with mobile devices. In *MobileHCI '17*, pp. 1–13. ACM Press, New York, New York, USA, 2017.

[9] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, m. c. schraefel, and J. O. Wobbrock. Reducing legacy bias in gesture elicitation studies. *interactions*, 21(3), 2014.

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

[11] T. Pham, J. Vermeulen, A. Tang, and L. MacDonald Vermeulen. Scale Impacts Elicited Gestures for Manipulating Holograms. In *the 2018*, pp. 227–240. ACM Press, New York, New York, USA, 2018.

[12] T. Pham, J. Vermeulen, A. Tang, and L. MacDonald Vermeulen. Scale impacts elicited gestures for manipulating holograms: Implications for ar gesture design. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pp. 227–240. ACM, New York, NY, USA, 2018.

[13] I. B. Rodriguez and N. Marquardt. Gesture Elicitation Study on How to Opt-in & Opt-out from Interactions with Public Displays. In *the Interactive Surfaces and Spaces*, pp. 32–41. ACM Press, New York, New York, USA, 2017.

[14] J. Ruiz and D. Vogel. Soft-Constraints to Reduce Legacy and Performance Bias to Elicit Whole-body Gestures with Low Arm Fatigue. In *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[15] C. Silpasuwanchai and X. Ren. Designing concurrent full-body gestures for intense gameplay. *International Journal of Human-Computer Studies*, 80:1–13, Aug. 2015.

[16] T. Tsandilas. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Trans. Comput.-Hum. Interact.*, 25(3):18:1–18:49, June 2018.

[17] R.-D. Vatavu and J. O. Wobbrock. Formalizing Agreement Analysis for Elicitation Studies. In *the 33rd Annual ACM Conference*, pp. 1325–1334. ACM Press, New York, New York, USA, 2015.

[18] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009.