Multimodal Technologies and Interaction

MDPI

Article

The Cost of Production in Elicitation Studies and the Legacy Bias-Consensus Trade off

Adam S. Williams [1,*], Jason Garcia [1], Fernando De Zayas [2], Fidel Hernandez [2], Julia Sharp [3] and Francisco R. Ortega [1,*]

1 Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA; J.S.Garcia@colostate.edu
2 Department of Computer Science, Florida International University, Miami, FL 33199, USA; fdeza001@fiu.edu (F.D.Z.); fhern103@fiu.edu (F.H.)
3 Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA; Julia.Sharp@colostate.edu
* Correspondence: AdamWil@colostate.edu (A.S.W.); F.Ortega@colostate.edu (F.R.O.); Tel.: +1-(970)-491-7445 (F.R.O.)

Received: 28 October 2020; Accepted: 2 December 2020; Published: 4 December 2020

Abstract: Gesture elicitation studies are a popular means of gaining valuable insights into how users interact with novel input devices. One of the problems elicitation faces is that of legacy bias, when elicited interactions are biased by prior technologies use. In response, methodologies have been introduced to reduce legacy bias. This is the first study that formally examines the production method of reducing legacy bias (i.e., repeated proposals for a single referent). This is done through a between-subject study that had 27 participants per group (control and production) with 17 referents placed in a virtual environment using a head-mounted display. This study found that over a range of referents, legacy bias was not significantly reduced over production trials. Instead, production reduced participant consensus on proposals. However, in the set of referents that elicited the most legacy biased proposals, production was an effective means of reducing legacy bias, with an overall reduction of 11.93% for the chance of eliciting a legacy bias proposal.

Keywords: legacy bias; elicitation; gesture; Wizard of Oz; methodology

1. Introduction

Gesture elicitation has become a popular study design that can be used to gain an understanding of interactions for emerging technologies and user behavior [1]. Through the use of elicitation methodologies paired with Wizard-of-Oz (WoZ) enabled systems, researchers can observe users' interactions with systems before accurate input recognition exists. Elicitation study design was popularized by Wobbrock et al. [2] in 2005. That design was further tested by members of the same team in 2009 [3]. This study methodology is under continual change and often improvement. Work has improved the metrics of consensus for interaction proposals [4,5]. More recently, metrics to assess the dissimilarity of gestures [6], and ways to identify the level of chance agreement in the study were added [7].

In an elicitation study, a participant is presented with a series of commands (referents) to execute through the use of gestures. Each referent is presented individually, often by text or animation [1]. For each referent, the participant proposes an interaction that would execute that command. In gesture elicitation studies these will be gesture proposals for multi-touch devices [3,8], or mid-air gestures [9]. Other modalities such as gesture and speech [10–13] can be elicited. Upon the generation of a proposal, the experimenter will trigger the reaction of the system to that proposal. This is called using a Wizard of Oz design. WoZ design allows the participant to feel that their interactions are actively being recognized, which may improve elicitation results, or at the very least, participant immersion.

A major criticism that elicitation studies currently face is the impact of prior technologies use on gesture proposals [14]. This bias is termed "legacy bias." As a result of this, studies have begun to explore ways to reduce legacy bias. Legacy bias could come from previous use of a device, an application, or a type of interaction. An open question in elicitation study methodology design is how to properly reduce legacy bias [14]. Legacy bias can be considered a desirable trait and has been leveraged to improve interaction design [8,15]. More often ways to reduce it have been attempted, feeling that a legacy biased interaction may not appropriately utilize the capabilities of a new system [16,17].

As of today, little to no work has examined legacy bias reduction techniques' impacts on the counts of legacy biased gestures. Instead, most studies use a legacy bias reduction technique as part of their methodology, but consider the resulting proposals from an input mapping and interaction design standpoint. Having participants produce more than one gesture per referent (production), and influencing a participants mindset before eliciting gestures (priming), are the most common forms of legacy bias reduction [1]. This research looks into whether the production technique reduces legacy biased gesture proposal frequency in elicitation studies. Some work has indicated that production may not produce the desired effect [8]. This is the first study that addresses this fundamental question in elicitation.

1.1. Motivation

Spatial interactive systems should be intuitive, discoverable, and easy to learn [2,18]. Elicitation as a form of participatory design can lead to the generation of an interaction set that exhibits those features. Participants are often videotaped while they generate proposals in responses to referents during and elicitation study [8,9,11,19]. Those videos are later analyzed based on the gestures used to generate a dataset. Video annotation has traditionally been done by hand [9,10]; however, some recent work has used skeletal data and computer vision [6]. The annotated gestures are then binned into equivalence classes based on their similarity [1,3]. After this binning, the agreement between participants' proposals is measured. At a high level, agreement metrics record the number of participants that agree on a given gesture for a given referent.

Gesture elicitation has been found to create guessable and intuitive input sets [20]. There is a growing body of evidence showing that users prefer user derived input sets over expert-defined ones [3]. That evidence helps to fuel elicitation methodology's increase in popularity. Out of this popularity, elicitation has seen use across a variety of domains including: multi-touch [9], internet of things [21], augmented reality [10,11,19], and interactive rings [22].

As elicitation studies are becoming increasingly prevalent, the question of legacy bias becomes an important one [16,17]. Care must be taken to ensure that the interactions derived from elicitation studies are not inappropriate for new domains. The recommendations from elicitation studies are being implemented [23,24], yet the impact of legacy bias and the effect of legacy bias reduction techniques have not yet been fully explored.

A study done on medical students and experienced anesthesiologists summarizes the motivation for this work [25]. This study elicited mid-air gestures for operating the anesthesia machines. The use of mid-air gestures in the operation room would reduce the risk of spreading diseases through touch. At the same time, if the gestures are not appropriate, the fine line between just enough and a deadly dose of anesthetic could be crossed.

The study found that the gestures produced by experienced practitioners were heavily influenced (biased) by the interactions with the current controls of the devices in question [25]. The produced gestures would imitate turning the knobs and flipping the switches presently found on the machine. In contrast, the medical students, who had far less exposure, generated more novel gestures [25]. While the study did not assess the merit of the relative gesture sets against each other, it may be the case that the novice gestures were more discoverable to other novices [3], or the novice gestures were more memorable [20]. Our concern is that the expert derived gestures were ill-suited for mid-air use,

where more optimal gestures could be used. Consider turning an invisible knob. The total space of available motion is small being limited by the wrist's range of motion, implying that the control display ratio would need to be high. If instead a gesture that imitated sliding an invisible slider was used the limiting factor for motion potential would be the shoulder and elbow instead of the wrist. This could allow a control display ratio to be below one. This lower ratio would allow fine-grained control over the flow of anesthesia.

1.2. Contribution

The major contributions of this research are:

- The first formal study that shows how production affects the frequency of legacy biased gesture proposals.
- An analysis of the trade-offs between legacy bias reduction and interaction set consensus.
- An examination of the referents where production works to reduce the frequency of legacy biased proposals.

2. Related Work

The authors of a 2012 study on multimodal elicitation found that some portion of participants interaction proposals were informed by interactions with prior technologies [12]. This finding was further fleshed out in their 2014 magazine article coining the term "legacy bias". That same article suggested three techniques for reducing legacy bias [14]. Since then, many elicitation studies have found that some non-trivial number of participants propose interactions that are informed by or exactly like interactions found with previous devices. This has been seen as a cut gesture being a finger imitation of scissors [19], zooming gestures that mirror what is used on smart phones [8,16,26], and even saying "F5" when prompted to produce a speech command for refreshing a web page [12].

An example of a legacy biased gesture proposal is using a "two-finger pinch" as seen in touch screens with a mid-air gesture system (left side of Figure 1). The gesture system could accept two hand expansions (right side of Figure 1) because of its extended recognition capabilities. In this case, either gesture could be appropriate. However, in other cases, the gestures may be limiting compared to the system's actual recognition capabilities. The methods suggested for reducing legacy bias are partnered elicitation, priming, and production [14]. Each suggested method has been used in elicitation studies [16,17,26,27]. However, this use was without an analysis of the impact of that reduction technique on the frequency of legacy biased proposals.

Priming is administering some sort of ques or activity to a participant before eliciting a gesture. An example is having participants do large range body motions like jumping jacks before producing a full-body mid-air gesture may cause them to generate a more physically involved gesture [14]. Priming has been done in a variety of ways, including, weighted constraints [16] and having participants do kinesthetic activities such as jumping jacks [17,26]. Often priming is done by more subtle means, seen as a description of the intended proposal space [28], or a visual and writing task used to add context to the elicitation study [29]. Production is asking participants to generate more than one gesture proposal, as seen in this study. The thought is that if the first gesture produced is legacy biased the next would not be. Paired elicitation is placing users in pairs or small groups and having them work together to generate new gestures by playing a variation of charades. Elicitation studies have been run with paired participants; however, this was not done to reduce legacy bias [12,27].

Production is the most frequently used technique used for legacy bias reduction [30]. In production studies, participants produce more than one interaction proposal per referent. Production commonly asks for three or more gestures [28,31,32]. Multiple legacy bias reduction techniques can be used in conjunction, seen commonly as a combination of priming and production [26,28].

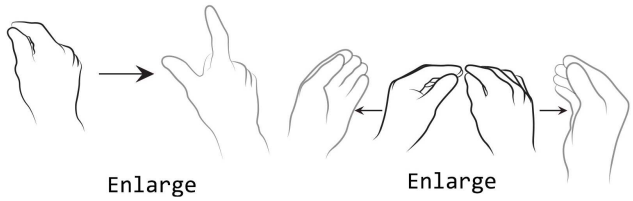

Figure 1. Legacy bias example (**left**): legacy biased, (**right**): new technique, Image used with permission from [33].

Legacy Bias Reduction

Very few studies have examined whether or not these techniques reduce legacy biased gesture proposals. One such study examined the impacts of production with priming on legacy biased proposal generation [17]. In that work the authors used 30 participants where all participants proposed 3 interactions each for 14 referents. Half of the participants were additionally primed by doing kinesthetic movements. The authors found that the effectiveness of these techniques may be minimal. More specifically they found that production had little practical effect while kinesthetic priming had a small effect [17].

Another work found that priming with information was ineffective however priming with physical constraints was more effective [16]. Physical constraints led to gesture proposals that used far smaller ranges of motion when compared to the unconstrained control group [16].

This work extends the limited work on examining the actual impacts of production as a legacy bias reduction technique by testing production alone, as compared to testing it in conjunction with priming [16,17]. In production studies, referent 2 proposal 1 is the 4th gesture a participant would have proposed not the 2nd, referent 1 elicited proposals 1–3. This means that the base frequency of legacy biased proposals cannot be estimated from the production trials due to the potential for the exhaustion of gesture proposals. This work is able to expand those previous works by utilization of a control group, thus allowing comparison of the control group's legacy biased proposal frequencies against the production groups. This work is the first paper to calculate the frequencies of elicited legacy biased proposals with a control group.

3. Materials and Methods

The study environment was designed in Unreal Engine 4 (Epic Games, Cary, NC, USA) using available assets from the unreal store. The environment was created to provide a sense of direction (ground, sky, etc.) for the participants as opposed to studies that used formless environments [8]. The system used a WoZ approach where the experimenter controls the movements of the environment as soon as the participant moved. The participants were told that the system would be able to recognize any gestures. This study was run on an HTC Vive HMD (HTC, Xindian, New Taipei City, Taiwan). Participants were recorded (externally) with an Xbox Kinect (Xbox Game Studios, Redmond, WA, USA) and a GoPro (GoPro, San Mateo, CA, USA). The display was recorded using Open Broadcaster Software (Open Broadcaster Software (OBS), https://obsproject.com). The GoPro footage was used to classify gestures. The participant's view from inside the headset was also recorded.

This experiment was a between-subject study with two groups: a control group, where the subjects were asked to provide one gesture per referent, and a production group where participants were asked to provide three gestures per referent. The referents were always presented randomly in both groups. Once the first proposal was elicited in the production group the participant was asked to provide two more gestures. The participants produced gestures for 17 different referents, listed in Table 1. These referents were selected to be realistic to travel and selection tasks in three dimensional (3D) immersive environments.

Table 1. Referents by category.

Translations	Rotations	Selection	Abstract
Move Up	Pitch Up	Select All Buttons	Destroy Green Button
Move Down	Pitch Down	Select Red Button	Duplicate Green Button
Move Left	Yaw Left	Select Red Buttons Only	
Move Right	Yaw Right		
Move Forward	Roll CW		
Move Backward	Roll CCW		

A total of 66 participants were recruited to take part in the experiment. Due to technical problems (e.g., data not recorded properly), only 54 participants (27 for the control group and 27 for the production group) were considered in the analysis. Participants' ages ranged from 18 to 31 with a median age of 22 years (21.96 control, 22.42 production), with 23 females and 31 males. The breakdown of the gender numbers for each group is listed in Table 2. No gender other than female and male were reported but the option was provided (e.g., non-binary, etc.). Thirty-six participants reported previously using a virtual reality device (17 control, 19 production) with 16 of them having direct experience with the HTC Vive (7 control, 10 production).

Table 2. Participants.

	Control	Production	Total
Male	19	12	31
Female	8	15	23
Total	27	27	54

Procedure

All subjects gave their informed consent for inclusion before they participated in the study. At the beginning of each session, participants were asked to complete an entry questionnaire which inquired about demographics, previous VR experience, and gaming habits. Once completed, participants were fitted with the HTC Vive headset and allowed to make adjustments for comfort. Before the experiment proper began, participants were placed in a standard prefabricated VR loading room and received a short training session to familiarize themselves with the elicitation procedure. The training session consisted of three simple tasks. Participants were asked to propose gestures that would create a sphere, cone, and cube. After they performed a gesture, the appropriate object was rendered and displayed in front of them.

Once the training session was completed, participants were presented with a short tutorial explaining pitch, yaw, and roll. These rotations were presented visually using a virtual model airplane shown from a 3rd person point of view. An image of the rotation tutorial, as seen by the participant, can be found in Figure 2. The model airplane was animated to perform each rotation in both directions. Participants were asked to verbally confirm that they understood the rotational definitions before proceeding.

Upon completing the tutorial, participants were placed in the custom VR environment which was designed to mimic a city street (Figure 3). Participants were instructed that they would be asked to propose gestures for each referent read aloud to them. Participants in the control group were told to propose a single gesture while the production group was instructed to produce three gestures for every referent. The order of the referents presented to each participant was randomized. After participants completed the elicitation experiment, they were given an exit survey.

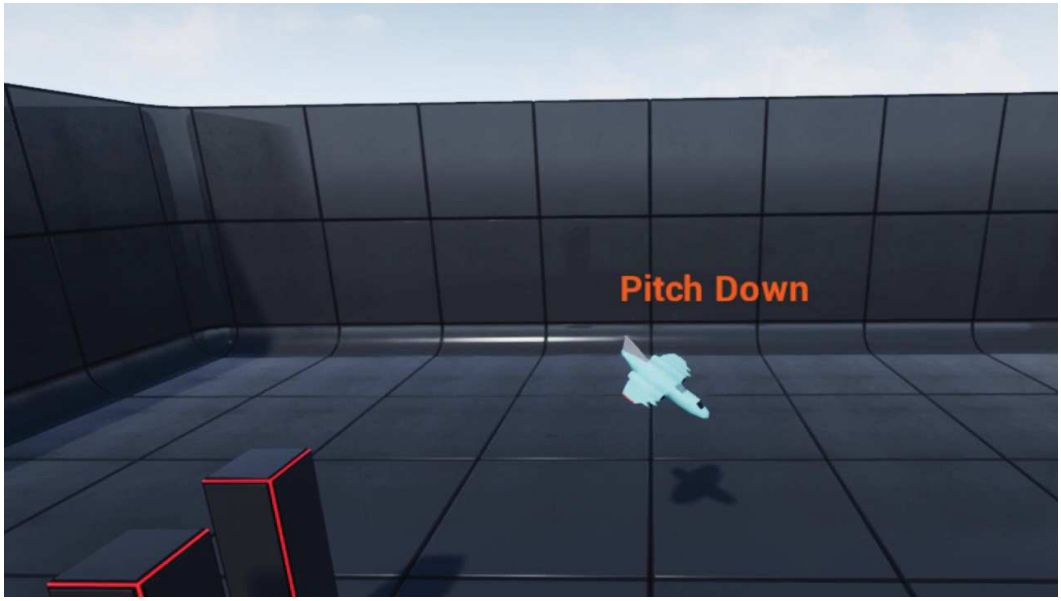

Figure 2. A virtual airplane from the rotational tutorial displaying an example of Pitch Down.

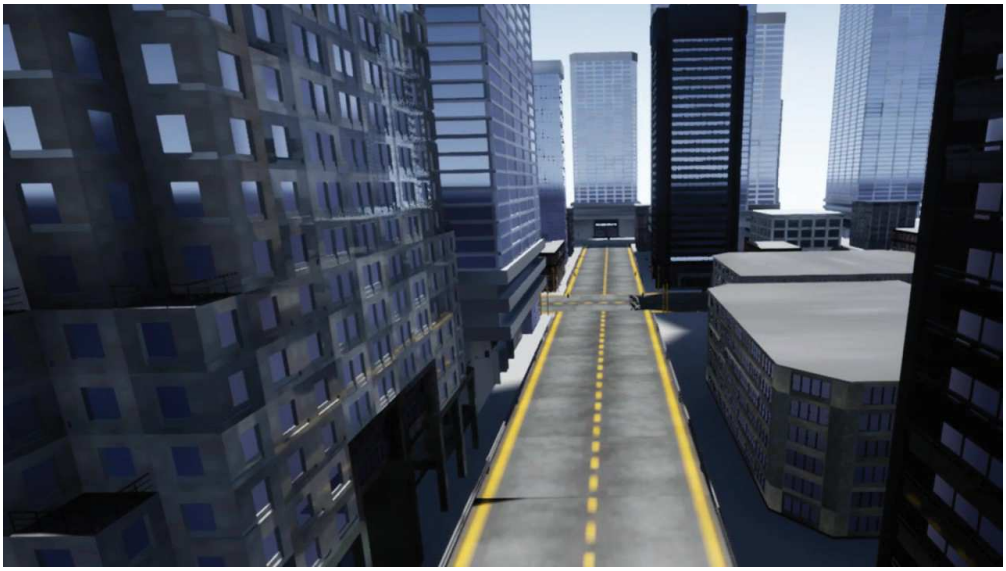

Figure 3. Illustration of the Elicitation Environment.

4. Results

The proposed gestures from both the control and production groups were reviewed and analyzed using the external video recordings from the GoPro. Gesture proposals were given descriptive identifying labels and gesture identifier numbers. An expert rater who was knowledgeable on gesture production then classified the gesture proposals into bins of equivalent gestures. Gestures were binned based on flow, nature, fingers used, palm shape, and the number of hands used. Flow was broken down into either discrete or continuous movements and nature was either physical or metaphorical. Physical gestures were ones that were physical representations of an action (e.g., direct manipulations) where metaphorical gestures were indirect interactions. Palm positions were coded based on orientation (i.e., facing forward, facing up). Previous work has shown that participants often alter the count of fingers used in similar gestures [26]. With that in mind, a one-finger swipe and a two-finger swipe were considered equivalent, while a swipe performed with both hands would fall into a separate

gesture class. Swipes with different orientations were considered distinct gestures based on the axis of motion. A swipe moving from left to right along the x-axis would be different from one moving down to up on the y-axis. No distinction was made between the specific direction on the axis used, a swipe left to right and one right to left was considered equivalent.

In total, 1836 gesture proposals were recorded and binned into 90 equivalence classes. These 90 gesture classes were analyzed for both legacy bias and consensus among participants. The effect of repeated gesture proposals on the presence of legacy bias was investigated, as well as the influence of specific referents on the likelihood of eliciting a legacy proposal.

To analyze the effects of legacy bias, the proposed gesture classes were categorized as either legacy or non-legacy gestures by a consensus of the independent votes of three expert raters. The If the gesture class could be identified with a known device in common usage, the proposal was classified as a legacy gesture. For example, a one-finger swipe would be classified as a legacy gesture, since it is a common interaction technique employed on smartphones and tablets. The gesture class Volume Knob, where a participant pantomimed the turning of a knob, was also classified as a legacy gesture since it can be identified with many common devices, say, a stereo for example. The list of legacy gestures classes found in our user study is given in Table 3.

Table 3. Gesture classes considered to be legacy.

Swipes	Taps	Other
One Hand Swipe X Axis (fingers)	One Hand Point and Tap	Volume knob
One Hand Swipe Y Axis (fingers)	Tap and Slide	
One Hand Swipe X Axis (hand)		
One Hand Swipe Y Axis (hand)		

When interpreting these results, keep in mind that production trial 1 is quite different than the control trial. On the first referent in the control group, the first gesture is proposed, on the second referent, the second gesture is proposed. This pattern continues for all referents. In the production group, the first referent elicits proposals 1–3 and the second referent elicits proposals 4–6. While the first gesture for each production trial should loosely follow the control gesture, the production group has potentially exhausted more gestures than the control group when reaching any referent apart from the first.

4.1. Effect of Referent on Legacy Bias

The raw count of legacy proposals for each referent was tallied overall four trials to examine the effect of an individual referent on the presence of legacy bias. The legacy counts for each referent can be found in Figure 4. The three selection tasks elicited the most legacy proposals. These referents were solely dominated by the mid-air One Hand Point and Tap legacy class. The referent with the highest count was Select Red Buttons Only with 84 legacy proposals out of possible 108. The single selection referent Select Red Button had 79 legacy proposals and the multiple selection task Select All Buttons had 65 legacy proposals. Among the selection referents, production does appear to reduce the frequency of legacy biased gesture proposals. The second production trial had 19.05% fewer legacy gestures proposed than the first, and the third trial has 15.69% fewer legacy proposals than the second.

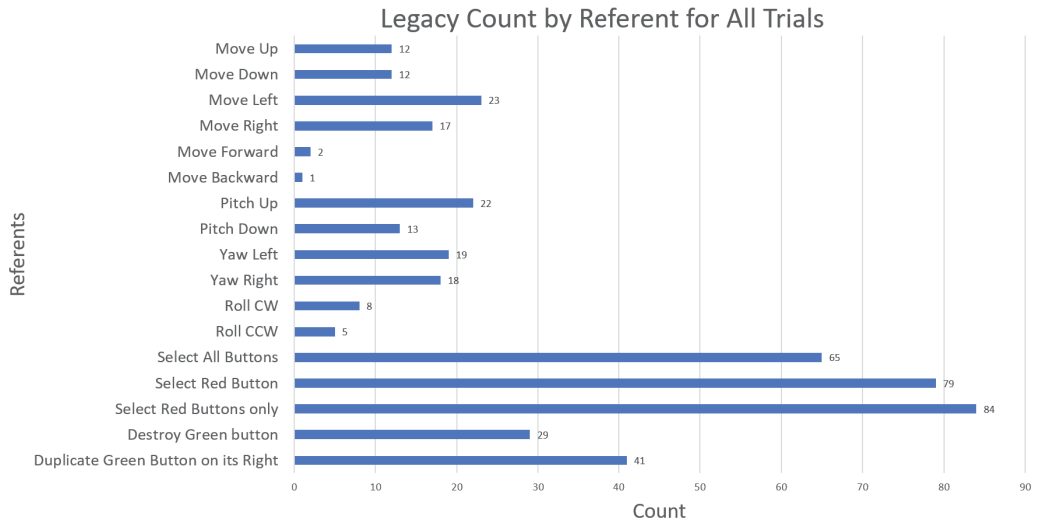

Figure 4. Legacy totals by referent over all trials. Each referent had 108 total gesture proposals.

4.2. Effect of Trial on Legacy Bias

In a preliminary analysis, the raw counts of legacy proposals were computed for each trial. The control group had 133 legacy proposals while the production group's first trial had 107 legacy proposals. With each trial consisting of 459 gesture proposals, this represents a difference of approximately 6%. Production's second trial had 115 legacy proposals, an increase of 8 legacy gestures over production trial 1. The lowest count occurred in the production group's final trial with 95 legacy proposals. The legacy counts and corresponding percentages for the control group and the three production trials are reported in Table 4.

Table 4. Legacy counts per trial. Each trial consisted of 459 proposals.

Group	Legacy Count	Percentage	Selection Referents Contribution to Total Legacy Count
Control	133/459	29%	71/133
Production Trial 1	107/459	23%	63/107
Production Trial 2	115/459	25%	51/115
Production Trial 3	95/459	21%	43/95

Production reduced the frequency of legacy-based gestures in the select referents which exhibited the highest legacy counts (Table 4). The difference between the control group and the production group's first proposal was eight (71 control, 63 production). This was about as large as the difference between production trial 2 and trial 3. It was also 2/3 of the difference between production trial 1, and production trial 2. The trend of decreasing each round gives evidence that production worked for those referents. The difference from the control group to the last production trial was 28 (71 control, 43 production trial 3).

4.3. Agreement Analysis

The proposed gestures were analyzed using the Agreement Rate formula, a gesture elicitation metric introduced by Vatavu and Wobbrock [4]. It provides a quantitative measure of participant agreement. For a single referent r, the Agreement Rate $AR(r)$ is defined as the number of participant pairs that were in agreement, divided by the total number of possible participant pairs. Two participants are said to agree if their gesture proposals are members of the same equivalence class. For example, we considered a one-finger swipe and a two-finger swipe (along the same axis)

as being equivalent. Therefore, the two participants proposing these gestures would be counted as a participant pair that are in agreement for a referent r. Formally, for a referent r, the Agreement Rate $AR(r)$ is given by Equation (1). Where P is the set of all proposals for referent r and the P_i are subsets of equivalent proposals from P. The Agreement Rates for all 17 referents over both the control and production groups can be found in Figure 5. Equation (2) was used to find the impact of legacy bias on Agreement Rate. In Equation (2) P is the set of all proposals for referent r and the L_i are subsets of equivalent proposals from P judged to be legacy gestures. The largest possible value for $LAR(r)$ is the original agreement rate $AR(r)$ for any given referent. The contribution of legacy gestures being agreed upon to the total agreement rate by referent is visualized by overlaying the legacy contribution on top of the total agreement rate in Figure 5. The agreement rate quantifies participant consensus by producing a value between 0 and 1. As a point of reference, agreement rates of 0.3–0.5 can be considered as high agreement for a sample size of 20 [4]. This study used a sample size of 54 allowing a rate of approximately 0.23 to be considered high.

$$AR(r) = \sum_{P_i \subseteq P} \frac{|P_i|(|P_i| - 1)}{|P|(|P| - 1)} \tag{1}$$

$$LAR(r) = \sum_{L_i \subseteq P} \frac{|L_i|(|L_i| - 1)}{|P|(|P| - 1)} \tag{2}$$

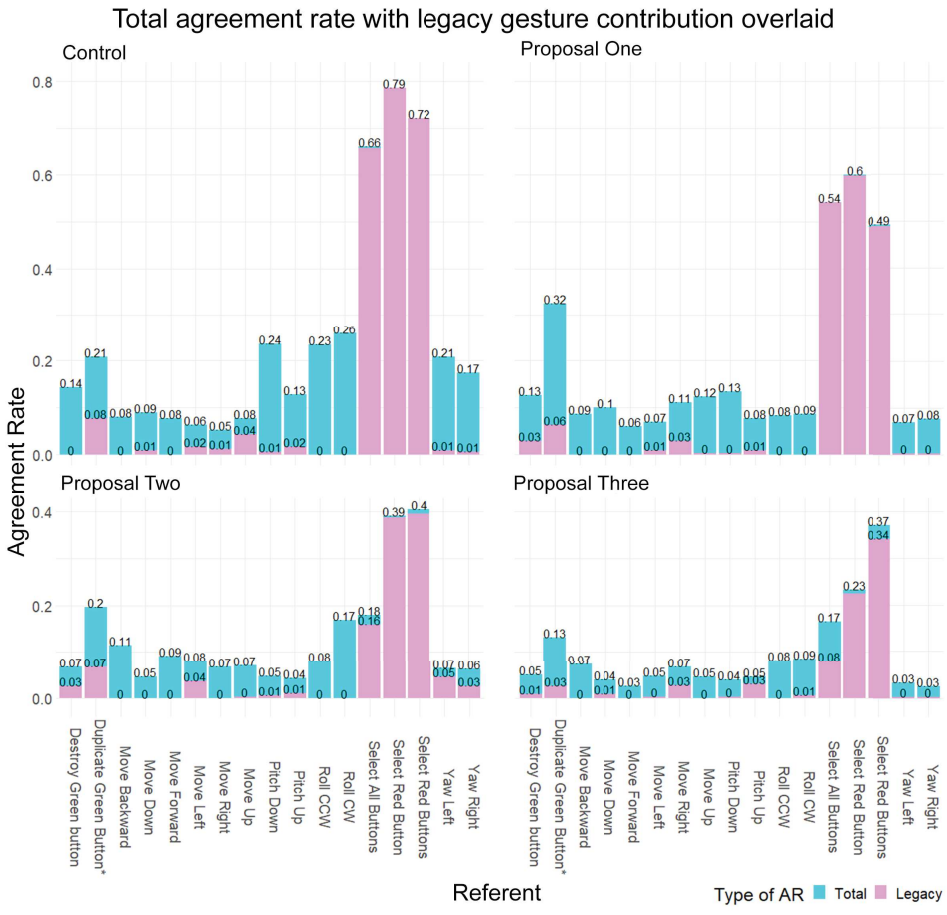

Figure 5. Agreement rates per referent for control and production groups with the contribution of agreement on legacy biased gestures overlaid. Legend: *: "Duplicate Green Button on it's Right", Charts have the same scale y-axis and the same x-axis.

Figure 5 shows that production causes fluctuations in Agreement Rates with a trend to decrease participant agreement over time. The most highly agreed upon gestures have most of their total agreement rate emerging from agreement on legacy gesture proposals. Over production trials, the agreement for these referents decreases while the contribution of legacy remains consistent. This is most visible for the selection-based referents which often had the touch-screen finger tap selection method used. Most of the travel referents achieved an agreement rate of less than 0.09. We speculate that part of the low agreement for these referents was due to travel in VR using gestures being relatively uncommon.

4.4. Legacy Proposal Frequency Trends for All Referents

These analyses attempt to find the impact of the condition (control, production trial 1, 2, 3) on the likelihood of proposing a legacy biased gesture. A chi-square test of independence showed that there was a significant association between condition and legacy proposal frequency $X^2\ (3,\ N\ =\ 459)\ =\ 8.9842, p = 0.03$. Based on this difference a chi-square test for trend in proportions was run and indicated that there is a trend in the proportion of legacy biased proposals dependent on condition ($X^2(1, N\ =\ 459)\ =\ 6.6151, p\ =\ 0.01$). A binomial logistic regression test with a logit link was used to further examine this trend. This test can be used to make a predictive model of a binary outcome variable with a multi-level explanatory variable. The results are listed in Table 5, and the resulting equation is shown in Equation (3). This analysis uses a dummy variable for the condition. The intercept is the base likelihood of producing a legacy gesture. The dummy variable for condition can be set to 0 or have one of the three trials set to 1, indicating which condition the prediction is for. Using this we can determine the probability of a legacy gesture for control only (all trials set to 0) or for any of the trials (that trial set to 1). The resulting probabilities are shown in Table 6. These probabilities indicate that production may help but that effect might be minimal and fluctuate depending on condition. Production trial 1 was 3.56% less likely than the control group to have a legacy gesture proposed. Production trial 2 was more likely to have a legacy biased proposal than production trial 1 and production trial 3 had the lowest chance of eliciting a legacy proposal.

$$Legacy = \beta_0 + \beta_1 Trial1 + \beta_2 Trial2 + \beta_3 Trial3 \quad (3)$$

Table 5. Logistical Coefficients for Legacy ∼ Trial.

Group	β_i	Std. Error	Z Value	$p(>\|z\|)$
Control (Intercept)	−1.2387	0.09848	−12.579	$<2 \times 10^{-16}$ ***
Production Trial 1	−0.2175	0.14568	−1.493	0.1354
Production Trial 2	−0.1454	0.14343	−1.014	0.3106
Production Trial 3	−0.3365	0.14967	−2.248	0.025 *

p-values: *** 0.001, * 0.05.

Table 6. Probability of getting a legacy biased gesture proposal by condition.

Control	Production Trial 1	Production Trial 2	Production Trial 3
22.47%	18.91%	20%	17.14%

4.5. Legacy Proposal Frequency Trends for the Selection Referents Only

When viewing the legacy proposal counts by referent (Figure 4), and the contribution of legacy gestures to total agreement rate (Figure 5) it is clear that legacy biased proposals are clustered around a few referents. Under these observations, we determined it appropriate to run the same tests run on the entire set of referents on the subset of selection referents, which exhibit the highest rates of legacy biased gesture proposals. This interpretation now examines: does production reduce the likelihood

of eliciting legacy biased proposals for referents which exhibit a high tendency towards eliciting legacy gestures?

As with the entire proposal set, a chi-square test of independence showed a significant association between the condition and legacy proposal frequency for the selection referents alone ($X^2\ (3,\ N\ =\ 81)\ =\ 27.474, p\ <\ 0.001$). A chi-square test for trend in proportions confirmed that there is still a trend in the proportion of legacy biased proposals dependent on the condition for the selection referents ($X^2(1, N\ =\ 81)\ =\ 27.284, p\ <\ 0.001$). To interpret what that trend is a logistic regression analysis was done in the same way as with all referents using only the selection referents. The equation remains the same (Equation (3)) while the results and coefficients are changed. The Logistic regression results for the selection referents alone is shown in Table 7 and the associated probabilities of encountering a legacy proposal are shown in Table 8. The predictions of the logistic regression now show a clear trend towards production reducing legacy bias in high-legacy referents. This trend starts with the first production trial, indicating that the methodology used in production (i.e., referent 1 = gestures 1–3, referent 2 = gestures 4–6,...) can reduce legacy bias even in the first proposal for each referent, in the case of high-legacy referents.

Table 7. Logistical coefficients for legacy ~ condition for the selection referents.

Group	β_i	Std. Error	Z Value	$p(>\|z\|)$
Control (Intercept)	−0.1318	0.1626	−0.811	0.4176
Production Trial 1	−0.1195	0.2338	−0.511	0.6091
Production Trial 2	−0.3309	0.2416	−1.369	0.1709
Production Trial 3	−0.5015	0.2491	−2.013	0.0441 *

p-values: * 0.05.

Table 8. Probability of getting a legacy biased gesture proposal by condition for the Selection Referents.

Control	Production Trial 1	Production Trial 2	Production Trial 3
46.71%	43.75%	38.64%	34.78%

5. Discussion

The main objective of this user study was to investigate if the production method for legacy bias reduction was effective. Over the 17 referents used in the study, the results indicate that the occurrence of legacy proposals did alter significantly when participants were asked to propose multiple gestures. However, the amount of this reduction appears to be minimal. In the referents that had high occurrences of legacy biased gesture proposals, the reduction was more visible. This reduction occurs even among the first gesture proposed in the production condition. The total impact of this was a reduction from 46.71% chance to produce a legacy biased gesture to a 34.78% chance in the third gesture proposal of a production study. It makes intuitive sense that a legacy bias reduction would work best on referents likely to produce legacy biased gestures, such as the selection referents.

This paper does not aim to provide a set of consensus gestures for use in VR travel applications, however; we believe that the legacy biased proposals suggested by participants are well suited for use in this domain. Out of the 90 binned gesture proposals found, 7 were considered legacy biased. These were “one hand point and tap”, “tap and slide”, “turning a volume knob”, and a “one hand swipe” on the x and y axis with either fingers used, or the entire hand used (Table 3). Of these, the “one hand point and tap” accounted for 51.11% of the total legacy biased gestures. This gesture was used primarily for selection across the select button referents. These legacy biased proposals seem appropriate for mid-air use in VR environments. None of them use interaction techniques that are ergonomically inappropriate for VR use. That said the “tap and slide” gesture may suffer from lack of haptic feedback on “tap” portion of the interaction.

The most common non legacy biased gesture for button selection was to reach out and "grab" the button. In this specific case the impact of legacy bias was beneficial. Across selection referents the "one hand point and tap" selection was proposed 228 times where "grab" was proposed 28 times. Either gesture can work well in VR; however, the legacy tap gesture was far more commonly suggested indicating a better fit for this environment. A counter example is seen in the "duplicate green button" referent where the most common proposal was to "grab and drag" the button. This command is more appropriate than the legacy biased proposal "tap and slide" due to the frequency of suggestion (43:"grab and drag", 24:"tap and slide"), however; both interactions use an ergonomic form that is appropriate to VR environments so this choice is based on participant consensus and not the level of the proposals legacy bias.

The most prominent feature of the agreement rate is that participant consensus was highest among the three selection tasks and that most of that agreement was caused by legacy gestures. This clustering implies that the actual level of legacy bias in an elicitation derived interaction set might be focused on a few interactions, such as selection and zooming operations. This clustering would likely occur with any interactions that are commonplace on cellphones or other ubiquitous devices. The analysis of agreement rates is promising as well as somber. Production typically decreased the agreement rate over proposals. By the third proposal, the agreement rate was lower in all but two cases. The good news is that the reduction in total agreement rate in the select referents was caused by a reduced occurrence of legacy biased gestures, meaning production worked to reduce legacy bias in those referents.

The simple selection referent Select Button was predominately achieved by a "point and tap gesture". The multiple selection referents had a large proportion of point and tap proposals as well. Most participants repeated the "point and tap" gesture for each sub-selection. Some interesting variants to this were observed, some participants used multiple fingers on the same hand to perform a simultaneous selection. Other participants used a similar technique using both hands.

A large proportion of the referents used in this study were given as 3D travel tasks. They accounted for 12 of the 17 referents. As observed in the agreement rates for these referents, there was very little consensus among participants for these tasks. The wide variation in proposals may reflect the lack of common 3D navigation exposure or that mid-air gesture is not an immediately apparent means for 3D navigation.

Agreement rates were low for all 12 travel referents with participants producing a variety of dissimilar gestures. Although the agreement rates for this referent category were negligible, we did identify a subset of participants who performed similar gestures for the Yaw Left/Right, Pitch Up/Down, and Roll clockwise/counterclockwise referents. This gesture imitated the motion of the airplane animation used in the pitch yaw roll tutorial. This airplane gesture was the most common gesture for these referents. That considered, the agreement rate given this accidental priming was still quite low. Out of this finding, we would recommend limiting the use of any easily imitated animation when describing rotations. The motion of the airplane animation may have altered the gestures proposed to match its movement.

Legacy bias is not inherently good or bad, but it is a common feature of elicited interaction sets. When conducting elicitation studies, decisions of whether legacy biased proposals are beneficial due to user familiarity and knowledge transfer from prior devices, or detrimental due to poor ergonomics or fit with the elicited technologies are ultimately left up to the practitioners researching that domain. Production may be used by researchers if they deem legacy biased proposals to be detrimental to their use case. When used, it should be expected to be minimally effective at reducing legacy bias across the referents. Production can help reduce the instances of legacy biased proposals for referents that are highly likely to exhibit legacy bias. In this study those referents were selections. When deciding to use legacy bias reduction techniques it is possible that the non-biased proposals generated are less well suited to the domain being used than the transferred legacy biased proposals. This should be considered when developing the elicited consensus set.

The impact of production on legacy biased proposals was minimal outside of the referents which exhibit high likelihood of eliciting legacy biased proposals. In other elicitation studies this is expected to remain consistent. That said, the use of the VR environment in this study may cause any gestures found here to be specific to the domain of VR travel and selection interactions. Outside of VR travel and selection we expect that the "one hand point and tap" gesture would occur with some frequency in referents from other domains that require selection.

6. Future Work

Posthoc analysis suggests that referents that were likely to elicit a legacy gesture were more likely to have the frequency of legacy proposals reduced by multiple trials. While this finding makes sense more rigorous investigation is needed to assess its validity. We would suggest an elicitation study that uses a larger set of referents that are likely to produce legacy gestures. We found that selection tasks, specifically mid-air selection tasks in a virtual environment, were highly influenced by legacy bias. These high-legacy referents should include a wider range of referent categories than explored in this work. Selections, zooming, scale adjustments, and other interactions common on touchscreen devices may all have increased rates of legacy bias proposals. After these high-legacy referents are identified, they should be examined with the production methodology to explore whether legacy proposals diminish over trials.

In studies where legacy bias gestures are concerned the formula in Equation (2) could be used to assess the impact of legacy gesture agreement on overall agreement rates. This style of transparent reporting may improve the dissemination of future elicitation study results.

We hope to see more elicitation studies use targeted legacy bias reduction mechanisms. Where legacy bias reduction methods are used for sets of referents that commonly exhibit legacy biased proposals. This may be difficult for elicitation studies that use a domain that has common use as with Morris, 2012, and web-browsing [12]. The results of this study were impacted by the choice of an uncommon elicitation environment. Mid-air gesture for travel in VR is uncommon. The low rates of legacy bias gestures and low agreement rates may have been contributed to by that. This domain choice may become more common with the advent of immersive analytics and its related need to travel data-sets in VR [34].

Future studies should examine the differences in the usability and goodness-of-fit of legacy biased gestures compared to the alternative gestures proposed in a production elicitation study. This work focused on the impact of production in lessening the frequencies of legacy bias proposals; however, the proposals that were generated in-place of the legacy biased proposals may not have been as well suited to the referents as the legacy biased proposals were. This is seen in the selection referents which had a single highly agreed upon legacy gesture that was appropriate to this domain that was replaced by varying gestures with lower consensus among participants.

7. Conclusions

Over the 17 referents, the production methodology only minimally reduced the likelihood of eliciting a legacy proposal. In the second proposal per referent in the production group, the rate of legacy biased gestures was actually increased. That said, production did reduce legacy proposals for the subset of referents that exhibited high tendencies towards eliciting legacy biased proposals. This reduction came at the cost of lower agreement rates. This reduction and associated lower agreement rates were present even in the first proposal made for each referent in the production condition.

Future elicitation studies should determine the importance of reducing legacy proposals compared to the goal of deriving discoverable interaction sets. In cases where legacy biased proposals can be assumed to be inappropriate for the technologies or environments being elicited, this trade-off may be appropriate. We believe that this trade-offs justification is seen in medical elicitation studies [25],

where experienced doctor's proposals could be perceived as inappropriate to the new interaction environment. In other cases, this trade-off may be inappropriate.

In either case, researchers should acknowledge that the production methodology for reducing legacy biased gesture proposals in elicitation studies will impact the results of the study starting from the first proposal for each referent. This novel finding has implications in the field of elicitation extending beyond the scope of this work. It was thought that production would not impact the first proposal for each referent, and never has it been found that production would reduce overall agreement rates. These are important discoveries that can help guide design choices in future elicitation studies.

Author Contributions: A.S.W.: Led the paper, helped define the legacy gesture set, wrote most sections of the paper, and ran the statistical analysis on the results. Roles: Formal analysis, Visualization, Writing—original draft, and Writing—review & editing. J.G.: Wrote a portion of the results section and made the associated tables used. Roles: Formal analysis, Visualization, Writing—original draft. F.D.Z.: Co-designed and developed the experiment. Oversaw data collection. Roles: Conceptualization, Investigation, Software. F.H.: Labeled the data, hand-coded all the participant videos, and created the gesture classes. Roles: Data curation. J.S.: Gave statistical advice and guided the analysis of the results. Roles: Supervision, Formal analysis. F.R.O.: Designed the experiment, supervised the entire process, wrote the introduction, assisted with editing the manuscript, and provided key elements for the experiment. Roles: Conceptualization, Methodology, Project administration, Supervision, Writing—original draft, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation (NSF) awards NSF IIS-1948254, NSF CCRI-CISE 2016714, and NSF BCS-1928502.

Acknowledgments: Thank you to Joseph Medina, Cristina Villarroel, Arelys Alvarez, Vanesa Perez, and Seidan Jamides for helping us during this long experiment.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Villarreal-Narvaez, S.; Vanderdonckt, J.; Vatavu, R.D.; Wobbrock, J.A. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies. In Proceedings of the ACM International Conference on Designing Interactive Systems (DIS'20), Eindhoven, The Netherlands, 6–10 July 2020; ACM Press: Eindhoven, The Netherlands, 2020.
2. Wobbrock, J.O.; Aung, H.H.; Rothrock, B.; Myers, B.A. Maximizing the guessability of symbolic input. In Proceedings of the Extended Abstracts on Human Factors in Computing Systems, Portland, OR, USA, 2 April 2005.
3. Wobbrock, J.O.; Morris, M.R.; Wilson, A.D. User-defined Gestures for Surface Computing. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 4 April 2009; ACM: New York, NY, USA, 2009; pp. 1083–1092.
4. Vatavu, R.D.; Wobbrock, J.O. 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, Seoul, Korea, 18 April 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1325–1334, [CrossRef]
5. Vatavu, R.D.; Wobbrock, J.O. Between-Subjects Elicitation Studies: Formalization and Tool Support. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7 May 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 3390–3402, [CrossRef]
6. Vatavu, R.D. The Dissimilarity-Consensus Approach to Agreement Analysis in Gesture Elicitation Studies. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow Scotland, UK, 2 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–13, [CrossRef]
7. Tsandilas, T. Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *ACM Trans. Comput. Hum. Interact.* **2018**, *25*, 18. [CrossRef]
8. Ortega, F.R.; Galvan, A.; Tarre, K.; Barreto, A.; Rishe, N.; Bernal, J.; Balcazar, R.; Thomas, J. Gesture elicitation for 3D travel via multi-touch and mid-Air systems for procedurally generated pseudo-universe. In Proceedings of the 2017 IEEE Symposium on 3D User Interfaces (3DUI), Los Angeles, CA, USA, 18–19 March 2017; pp. 144–153.

9. Ortega, F.R.; Tarre, K.; Kress, M.; Williams, A.S.; Barreto, A.B.; Rishe, N.D. Selection and Manipulation Whole-Body Gesture Elicitation Study In Virtual Reality. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; pp. 1723–1728.
10. Williams, A.S.; Garcia, J.; Ortega, F. Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 3479–3489, [CrossRef] [PubMed]
11. Williams, A.S.; Garcia, J.; Ortega, F. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–21, [CrossRef]
12. Morris, M.R. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces, Cambridge, MA, USA, 11 November 2012; ACM: New York, NY, USA, 2012; pp. 95–104, [CrossRef]
13. Khan, S.; Tunçer, B. Gesture and speech elicitation for 3D CAD modeling in conceptual design. *Autom. Constr.* **2019**, *106*, 102847. [CrossRef]
14. Morris, M.R.; Danielescu, A.; Drucker, S.; Fisher, D.; Lee, B.; Schraefel, M.C.; Wobbrock, J.O. Reducing Legacy Bias in Gesture Elicitation Studies. *Interactions* **2014**, *21*, 40–45. [CrossRef]
15. Köpsel, A.; Bubalo, N. Benefiting from Legacy Bias. *Interactions* **2015**, *22*, 44–47, [CrossRef]
16. Ruiz, J.; Vogel, D. 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, Seoul, Korea, 18 April 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 3347–3350, [CrossRef]
17. Hoff, L.; Hornecker, E.; Bertel, S. 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, Eindhoven, The Netherlands, 14 February 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 86–91, [CrossRef]
18. Nielsen, M.; Störring, M.; Moeslund, T.B.; Granum, E. A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*; Camurri, A., Volpe, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 409–420.
19. Piumsomboon, T.; Clark, A.; Billinghurst, M.; Cockburn, A. User-Defined Gestures for Augmented Reality. In Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 955–960, [CrossRef]
20. Nacenta, M.A.; Kamber, Y.; Qiang, Y.; Kristensson, P.O. Memorability of pre-designed and user-defined gesture sets. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April 2013.
21. Zaiţi, I.A.; Pentiuc, Ş.G.; Vatavu, R.D. On free-hand TV control: Experimental results on user-elicited gestures with Leap Motion. *Pers. Ubiquit. Comput.* **2015**, *19*, 821–838. [CrossRef]
22. Gheran, B.F.; Vanderdonckt, J.; Vatavu, R.D. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In Proceedings of the 2018 Designing Interactive Systems Conference, Hong Kong, China, 8 June 2018; pp. 623–635, [CrossRef]
23. Huang, Y.J.; Fujiwara, T.; Lin, Y.X.; Lin, W.C.; Ma, K.L. A gesture system for graph visualization in virtual reality environments. In Proceedings of the 2017 IEEE Pacific Visualization Symposium (PacificVis), Seoul, Korea, 18–21 April 2017; pp. 41–45.
24. Piumsomboon, T.; Altimira, D.; Kim, H.; Clark, A.; Lee, G.; Billinghurst, M. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In Proceedings of the 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 10–12 September 2014; pp. 73–82.
25. Jurewicz, K.A.; Neyens, D.M.; Catchpole, K.; Reeves, S.T. Developing a 3d gestural interface for anesthesia-related human-computer interaction tasks using both experts and novices. *Hum. Factors* **2018**, *60*, 992–1007. [CrossRef] [PubMed]
26. Wittorf, M.L.; Jakobsen, M.R. Eliciting Mid-Air Gestures for Wall-Display Interaction. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction, Gothenburg, Sweden, 23 October 2016; pp. 3:1–3:4.

27. Nebeling, M.; Huber, A.; Ott, D.; Norrie, M.C. 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, Dresden, Germany, 6 November 2014; pp. 15–24, [CrossRef]
28. Chan, E.; Seyed, T.; Stuerzlinger, W.; Yang, X.D.; Maurer, F. User Elicitation on Single-Hand Microgestures. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7 May 2016; pp. 3403–3414, [CrossRef]
29. Cafaro, F.; Lyons, L.; Tarre, K.; Antle, A. 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, Montreal, QC, Canada, 21–26 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–12.
30. Vogiatzidakis, P.; Koutsabasis, P. Gesture Elicitation Studies for Mid-Air Interaction: A Review. *Multimodal Technol. Interact.* **2018**, *2*, 65. [CrossRef]
31. Lee, L.; Javed, Y.; Danilowicz, S.; Maher, M.L. Information at the Wave of Your Hand. In Proceedings of the HCI Korea, Seoul, Korea, 10 December 2014; pp. 63–70.
32. Koutsabasis, P.; Domouzis, C.K. Mid-Air Browsing and Selection in Image Collections. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Bari, Italy, 7 June 2016; pp. 21–27, [CrossRef]
33. Williams, A.S.; Ortega, F.R. Evolutionary Gestures: When a Gesture is Not Quite Legacy Biased. *Interactions* **2020**, *27*, 50–53, [CrossRef]
34. Marriott, K.; Chen, J.; Hlawatsch, M.; Itoh, T.; Nacenta, M.A.; Reina, G.; Stuerzlinger, W. Immersive analytics: Time to reconsider the value of 3d for information visualisation. In *Immersive Analytics*; Springer: Basel, Switzerland, 2018; pp. 25–55.