An Evaluation of Touchless Hand Gestural Interaction for Pointing Tasks with Preferred and Non-preferred Hands

Alvin Jude
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
alvin_jude@baylor.edu

G. Michael Poor
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
michael_poor@baylor.edu

Darren Guinness
Dept of Computer Science
Baylor University
Waco, TX 76798 USA
darren_guinness@baylor.edu

ABSTRACT
Performance evaluations of touchless gestural interaction are generally done by benchmarking pointing performance against existing interactive devices, requiring the use of user's preferred hand. However, as there is no reason for this interaction to be limited to only one hand, evaluation should rightfully consider both hands. In this paper we evaluate the performance of touchless gestural interaction for pointer manipulation with both the preferred and non-preferred hands. This interaction is benchmarked against the mouse and the touchpad with a multidirectional task. We compared the performance between all devices, improvement in performance between 2 rounds, and the degradation of performance between hands. The results show the mouse has no performance increase between rounds but high degradation across hands, the touchpad has medium performance increase and medium degradation, and gestural interaction has the highest performance increase and the lowest degradation between hands.

Author Keywords
Fitts' Evaluation; Non-Preferred Hand; Gestural Interaction; Motor Learning;

ACM Classification Keywords
H.5.2 User Interfaces: Evaluation/methodology, Input devices and strategies

INTRODUCTION
Touchless gestural interaction tends to be an effective, intuitive, and natural interaction for users to relay information to the computer [25]. Currently, commodity hardware such as the Leap Motion and Xbox Kinect is readily available to the general public. However, these devices are targeted towards non-work, non-purposeful interactions often used in isolation (such as gaming). This limited functionality is the exact opposite of the multi-functional concept of desktop computers [1]. While touchless gestural interaction can definitely be fun [16], there is no reason it cannot also be functional. It should be possible to incorporate gestures into everyday devices used for work-related, purposeful interactions, such as pointer manipulation, 2-handed desktop interaction, or as a secondary input device.

The multi-purpose use of gestures should result in a highly usable mode of interaction, as it employs well-researched principles of Reality-Based Interaction. It provides a natural and intuitive method of interaction, by building on users' pre-existing knowledge of the everyday, non-digital world [8]. Wigdor and Wixon state that natural user interfaces should should be designed to mimic some other experience at which users are already experts, thus allowing them to feel "like a natural". They also strongly suggest that natural user interfaces should provide an experience that is just as natural to a novice as it does to an expert user [27]. This gives us a theoretical foundation to believe that touchless gestural interaction should be a highly usable, natural mode of interaction. However, little is known on how effective it would be in the real world, it's strengths and limitations, nor how it would compare relative to existing devices. Therefore we sought to develop a better understanding of the capabilities and potential of this interaction by performing an evaluation of pointer manipulation tasks on both the preferred and non-preferred hands.

There is no reason to limit touchless gestural interaction to using only one hand. Instead, it could potentially be used in 2-handed interaction such as bi-modal interaction where both hands are used simultaneously, as an optional secondary (or tertiary) input device, or to allow users to alternate between hands based on preference or comfort levels. This formed the basis of our experiment, which was designed to benchmark gestural interaction using both hands against the mouse and the touchpad. We examined each device's improvement between rounds, as well as degradation or difference in performance between hands.

Our results show lower performance with gestures compared to the other devices, which is generally the case with new devices. However, gestures had the highest performance improvements between 2 rounds, and the lowest degradation between hands. We believe the lower degradation will lead to better use of the device on the non-preferred hand, thus paving way for the 2-handed interactions mentioned above.

We also observed interesting findings in the other devices used in our experiment. For example, the mouse showed no

NordiCHI '14, October 26–30 2014, Helsinki, Finland.

ACM 978-1-4503-2542-4/14/10...$15.00.
http://dx.doi.org/10.1145/2639189.2641207

significant performance increase on neither the dominant nor the non-dominant hand, leading us to believe that knowledge of the device is transferred between hands.

RELATED WORKS

Pointing Device Evaluation

In 1954 Paul Fitts established a relationship between movement speed and accuracy in rapid motor movements [5]. This relationship paved way to standardized pointing device evaluations used today. Fitts introduced what he initially termed Index of Performance, which has since been referred to as throughput and calculated as the index of difficulty (ID) over movement time (MT) [30],

$$\mathrm{Throughput} = \frac{\mathrm{ID}}{\mathrm{MT}} \tag{1}$$

where ID is the ratio of the distance(D) to width(W) and measured in bits:

$$\mathrm{ID} = \log_2\left(\frac{2\mathrm{D}}{\mathrm{W}}\right) \tag{2}$$

This definition has been recently updated [22] to use the Shannon formulation, so:

$$\mathrm{ID} = \log_2\left(\frac{\mathrm{D}}{\mathrm{W}} + 1\right) \tag{3}$$

The original definition of distance (D) in the formula above measures the distance between the center point of two targets. In implementations where the targets are circles, width (W) represents the diameter of the circular target. One such implementation is illustrated in Figure 2.

This formula has been adjusted in recent literature to use the distance between the starting point of the cursor to the ending point, known as effective distance ($\mathrm{D_e}$). Width has also been updated to use effective width ($\mathrm{W_e}$), which is 4.133σ where σ is the standard deviation of the actual endpoint distribution. Current methods of calculating throughput therefore uses an effective index of difficulty ($\mathrm{ID_e}$) [30, 22] defined as:

$$\mathrm{ID_e} = \log_2\left(\frac{\mathrm{D_e}}{\mathrm{W_e}} + 1\right) \tag{4}$$

This adjusted definition of effective index of difficulty is used in our throughput calculation. Therefore, we calculate throughput as

$$\mathrm{Throughput} = \frac{\mathrm{ID_e}}{\mathrm{MT}} \tag{5}$$

Handedness

Handedness refers to the performance of the preferred and non-preferred hands; there are several reasons to study the non-preferred hand in pointing tasks. Studies have shown that with increased use of the mouse, we are exposed to higher risk of hand impairments such as carpal tunnel [12]. 2-handed gestural interaction allows a potential solution by allowing the user to use a more natural mode of interaction such as gestures, and by allowing users to alternate between hands. A smaller difference in performance between hands could better encourage this behavior. Additionally multi-handed gestural interaction has demonstrated usefulness in areas such as robot control [29] and neurosurgery [6].

One of the first studies we found to examine handedness in human-computer interaction(HCI) reported differences in performance between hands, which they referred to as degradation [10], defined as:

$$\mathrm{Degradation} = \frac{\mathrm{Time_{NPH}} - \mathrm{Time_{PH}}}{\mathrm{Time_{NPH}}} \tag{6}$$

In degradation as per Equation 6, $\mathrm{Time_{NPH}}$ refers to the mean movement time of the non-preferred hand measured in milliseconds, while $\mathrm{Time_{PH}}$ refers to that of the preferred hand. The researchers mentioned above reported that the mouse and stylus degraded at about the same rate of 28% while the trackball did not degrade between hands. They also reported that similar performance was observed between hands for larger targets and larger distances.

Gestural Pointing

Pino et al looked into the use of the Xbox Kinect for point-select tasks [18]. The experiment utilized both a 2D and 3D multi-directional task to evaluate the pointing device. The 2D task used 5 blocks of 15 selection targets, resulting in 75 (5×15) trials per task. The experiment reported throughput of 2.10 bits/s for the 2D tasks, and 1.06 for the 3D tasks. The Kinect's throughput was 39% lower for the 2D task, but outperformed the mouse in the 3D task by 9.7%. The study also reported that Fitts' law extended into gestures similarly to the mouse in the 2D task and at a higher rate in the 3D task.

Sambrooks and Wilkinson looked into comparing touch, gestures, and mouse interactions [19]. This work focused on the performance of touch versus gestures, the relationship of gesture performances over time, and if gestural performance suffered from fatigue. The study used a task with 100 targets further broken into rounds of 20 targets, with 10 second breaks between rounds. Selection was done with a left-click for the mouse and single-tap for touch. For gestures, the study used an animated 'selection circle' when hovering over a selection element. The study found that fatigue was not a factor in gestures and that there was no improvement in gestures over the course of the experiment. We designed our study to take a deeper look into this claim, as we found it to be counter-intuitive. New interactions generally do exhibit performance improvements [15], and gestures being a more natural mode of interaction should too.

A previous study compared the mouse and two gestural interaction techniques and identified that the "Gorilla Arm Syndrome," a known fatigue problem of touch screen interaction, extended to gestural interaction [9]. The authors proposed a solution to this problem by allowing the user to perform gestural interaction from a resting position. This technique – referred to as the Personal Space approach – allows the user to build their interaction space by selecting 4 corners while their elbow is in a rested position. These four points selected during the calibration process is used to map the user's interaction space to the screen coordinates. The experiment noted that fatigue was mentioned within 90 seconds of use in the standard gestural pointing technique, while no fatigue was reported with their technique. The study also reported no significant difference in terms of movement time between both methods of gestural interaction. Our gestural interaction style was based on the approach introduced here.

Learning Effects

An evaluation of the mouse, trackball, joystick, and touchpad using a 2D multi-directional point-select task was performed in [15]. The experiment took the learning effects of the interface into consideration, and only performed analysis on the data where there was no learning present. In order to achieve this, each device was tested over 10 blocks each with 5 sequences of 15 target selections. Each device took about an hour to evaluate and each participant was evaluated with all devices. They found that learning effects were not significant after 5 blocks. In our experiment, we measure and compare the learning effects described here.

Motor Learning

As we were evaluating a novel interaction across both hands, we could not allow the user to perform the task until no performance improvement was observed. Therefore, we decided to investigate the learning effects of the device. Motor Learning, as defined by Schmidt and Lee, is described by four distinct characteristics [20]: (1) Learning is a process of acquiring the capability for producing skilled actions. (2) Learning occurs as a direct result of practice or experience. (3) Learning cannot be observed directly, as the processes leading to changes in behavior are internal, and are usually not available for direct examination. (4) Learning is assumed to produce relatively permanent changes in the capability for skilled behavior. Schmidt and Lee indicated that while we cannot directly observe learning, we can measure and report performance improvements, from which we can infer learning. In this study we measured and reported performance improvements between rounds, with the commonly used measure of effect size [14, 23, 21].

Lin et al [14] indicated that random-order practice (e.g. A-B-C, B-C-A, C-A-B) generally benefits motor learning more than block order practice (A-A-A, B-B-B, C-C-C). As we were evaluating performance improvement, it was in our best interest to optimize motor learning. We therefore used this method in designing our experiment, where each participant used 2 devices on both hands per block, and each block performed twice.

Random-order practice mentioned above and utilized in our experiment has been heavily studied outside of HCI, and is generally referred to as 'distributed practice', whereas the block order practice mentioned above is referred to as massed practice. Distributed practice has been shown to be better for performance improvement than massed practice [13].

METHODOLOGY

Participants

A total of 36 participants (M=22, F=14) were recruited to participate in the user study. Participants' age ranged from 18-36 years with a median of 21 years. Participants' computer use ranged from 1 to 50 hours per week with a median of 30 hours. All participants were undergraduate or graduate students who were compensated for their participation.

Participants were placed in either one of 3 groups, where each group would perform the given tasks with 2 different devices. The first group used gestures and mouse, the second used gestures and touchpad, the third used the touchpad and the mouse. This gave us 24 samples per device.

Task

The task performed by the participants consisted of 70 trials. In each trial, a square target appears at locations designed to look random. Participant were required to hover on the target for 500 milliseconds, which caused the target to disappear and a new target to appear. This action is repeated until the task ends.

Four task profiles were created, each with 70 trials. These trials were specifically designed to incorporate different direction and distance between trials. All 4 task profiles had 15 small, 20 medium, 20 large and 15 extra-large targets, which were respectively 90, 120, 160, and 220 pixels in both height and width. All 4 task profiles had a similar ID rating, ranging from 200 to 210 bits.

We aimed to make the experiment fun and to encourage speed. We therefore attempted to gamify the task by asking the users to complete as quickly as possible, and included a status cell that displayed score and time per task.

During our pilot study, we identified the need for high contrast between the cursor, targets, and the background. We therefore used a black background, with white targets, and a green cursor which was high in contrast to both the background and the targets. The cursor was a 32x32 pixel image, the largest cursor size allowed by Microsoft Windows 7. These changes helped participants identify their cursor easily on the screen, thus increasing the speed at which they could complete a trial. The larger size was in part necessary due to the larger monitor used (30 inches) which was positioned approximately 3 feet (90cm) from participants.

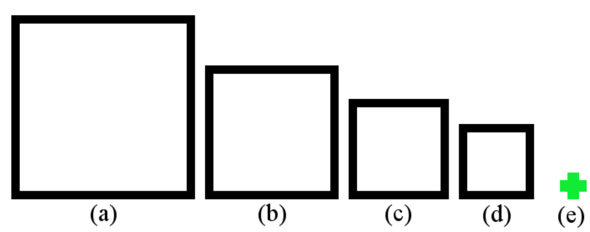

Figure 1. A scaled version of the targets used. (a) 220x220 pixels, (b) 160x160 pixels, (c) 120x120 pixels, (d) 90x90 pixels. The cursor shown in (e) was set to a high contrast green shown 32x32 pixels.

Breaking the Standard Evaluation Method

The ISO 9241-9 standard has outlined an effective evaluation method for pointing devices. The 2D multi-directional task (illustrated in Figure 2) states the following are necessary [22]: (1) Circular or square targets may be used, (2) The

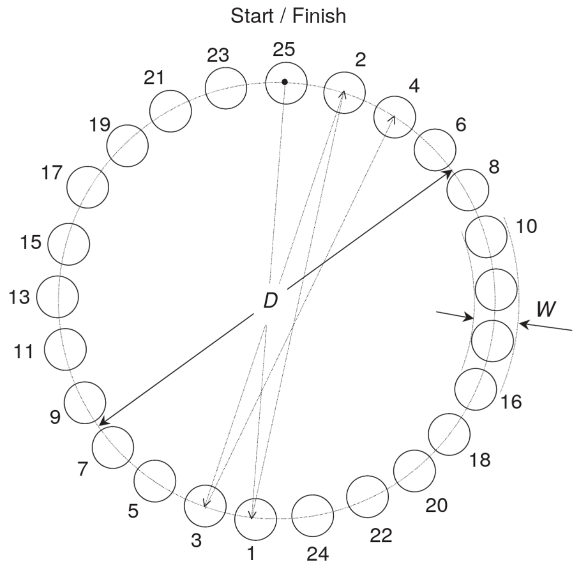

Figure 2. Multi-Directional task implementation from [22] used with permission. The distance (D) and width (W) can be plugged into Equation 3

effect of direction should be controlled, (3) The path should begin and end at the same target, and (4) The software must graphically indicate which target the participant should proceed to next.

This standard has been widely accepted by the HCI community and is highly recommended for pointing device evaluation [22]. However, we had to break away from this implementation as our pilot studies showed that the corners of the screen were the most difficult for users to reach with gestural interaction. Using the standard method would not allow us to assess this behavior, or to factor this behavior into the experiment, which would mean the results would be flawed, or based on a highly optimistic measure.

Other non-standard features were incorporated in our system. A hover action, similar to that used in another gestural study [19] was used instead of a select action. This was done as we did not want to introduce any nuisance variable by implementing gesture-based selection. The targets were made to appear at random locations, with no graphical indication where the next would be. This controlled for muscle memory and promoted stronger learning [7]. Finally, the same location was not chosen for both the beginning and end. This allowed us to reach more areas of the screen with the same amount of trials. We did however take 2 of the suggestions into consideration: we controlled for the effect of direction, and created trials with ID between the recommended range (1-4.5 bits).

Input Devices & Interaction

The devices used in our experiment were selected for their equally usable nature on both hands without being biased to either. We used the a symetrical mouse (Logitech M-U0032-O), a standalone external touchpad (PERIPAD-702), and the Leap Motion controller for gestures. Participants were encouraged to place the peripherals in a position that was most comfortable for them. Researchers helped adjust the angle and position of the Leap Motion controller to increase recognition rates.

The standard method of interaction for gestural input involves the user holding their arms out, which is known to cause fatigue [24, 26], commonly known as the "Gorilla Arm Syndrome" [2]. We addressed this issue by using the Personal Space approach [9] as it was demonstrated to heavily reduce fatigue without sacrificing performance, by allowing users to rest their elbow on a surface. The users were required to define their own interaction space in 3 dimensions through a calibration stage as shown in Figure 3.

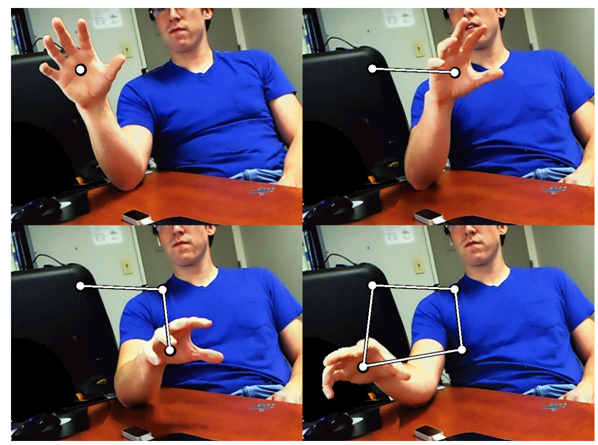

Figure 3. Personal Space calibration: Constructing the user's space

In this experiment, the software prompted participants to position their hand where they would like each screen corner to be. Most participants would anchor their screen boundaries by pointing to the corner of the screen with their hand. This created a more intuitive calibration, which allowed the participants to easily navigate in 3 dimensional space.

After the calibration, the participants were asked to navigate around the screen. The 4 corners were given specific attention as we found this to be the most sensitive regions for recognition. Most participants were able to find a comfortable calibration by their second attempt during round 1 of the experiment, and on their first attempt on the second round.

Procedure

The participants were first shown a 1-minute video detailing the hover task. Before using each device, they were shown a short video explaining the the device being used. The gestural interaction had a longer video of approximately 3 minutes, including an explanation of the calibration process.

Participants would perform the task once on each device on each hand, totalling to 4 tasks per round. Each task is done with a different task profile, therefore ensuring that the participants would not be able to learn the task or guess the position of each target. A practice session consisting of 12 trials was performed before the full task of 70 trials. The 4 tasks in the first round was followed by a 2 minute break, after which the participants were asked to perform the exact same tasks as before in the same order and on the same task profile. The training task was not present in the second round.

Upon completion, participants were given an exit survey detailing their prior use of each device and their preference, comfort and fatigue experienced with those devices during the experiment. The entire experiment lasted between 45 to 60 minutes.

Design

The experiment used a between groups design with 3 groups. A 2x2 latin-squares designed was used to counterbalance the order in which the device and hands are used in an experiment. We had the option of performing the tasks in orders across 2 rounds (A-B-C-D, A-B-C-D) or in blocks (A-A, B-B, C-C, D-D), and we chose the former as it has been shown to generally benefit motor learning [14].

RESULTS

Preferred Hand vs Non-Preferred Hand

All but 3 of our 36 participants self-reported as right-handed; 2 self-reported as left-handed while 1 self-reported as ambidextrous. Nonetheless, all 36 participants reported their right hand as the preferred hand for using both the mouse and the touchpad. This included the participants who did not report as right-handed. As a result, we refer to the right hand as the preferred hand and the left hand as the non-preferred hand in the analysis.

Time

Two time-related metrics are commonly used in literature: completion time and movement time [30]. Completion time is shown in Table 1. This metric provides good indication of the duration in which each participant took to complete each task, and a simple overview of the performance of each device. It is however not suitable for analysis as it includes the hover time of 500 milliseconds per trial or 35 seconds per task. Adding a constant factor to the metrics collected causes the differences between them to appear much smaller than the actual difference.

	Source	Mouse		Touchpad		Gestures	
		Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
R	Round 1	78.8	3.4	94.8	5.7	121.2	14.4
	Round 2	78.5	2.9	92.6	6.1	115.7	11.2
	Overall	78.4	3.1	93.7	6.0	118.5	13.1
L	Round 1	98.6	8.2	111.8	13.2	131.5	19.5
	Round 2	97.9	8.7	107.7	11.4	124.1	14.3
	Overall	98.2	8.4	109.7	12.4	127.8	17.3

Table 1. Means of completion time in seconds for each device across both rounds and for both the right (R) and left (L) hand.

Movement time refers to the time taken for the participant to move from the starting point to the target and therefore does not include the hover time of 500 milliseconds. Table 2 shows the mean movement time per trial across all devices and hands. This table is presented to allow for a comparison with existing literature such as that of Kabbash et al [10], but this metric has since been shown to be naive [30], and therefore not used for further analysis. We did however notice improvements in movement time between rounds on all devices across hands, with the exception of the mouse used with the right hand.

	Source	Mouse		Touchpad		Gestures	
		Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
R	Round 1	626	49	855	82	1232	206
	Round 2	621	41	823	88	1153	161
	Overall	623	45	839	85	1193	187
L	Round 1	909	117	1097	189	1379	278
	Round 2	898	125	1038	163	1274	204
	Overall	904	120	1068	177	1327	247

Table 2. Mean Movement Time per trial in milliseconds.

Throughput

We measure performance by throughput, defined as

$$\mathrm{Throughput} = \frac{\mathrm{ID_e}}{\mathrm{MT}} \tag{7}$$

The means of the throughput is shown in table 3 along with the standard deviation. An anova test performed on the pooled throughput between rounds shows a significant difference in the means $(F(5, 282) = 240.76, p < 0.001)$. A Tukey's HSD test shows statistically significant difference in means across all devices and hands. We can therefore rank the devices based on the overall means from best to worst as such: right mouse, right touchpad, left mouse, left touchpad, right gestures, left gestures.

	Source	Mouse		Touchpad		Gestures	
		Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
R	Round 1	4.81	0.42	3.58	0.34	2.56	0.33
	Round 2	4.79	0.38	3.74	0.47	2.71	0.30
	Overall	4.80	0.39	3.66	0.42	2.64	0.32
L	Round 1	3.36	0.38	2.86	0.44	2.33	0.36
	Round 2	3.40	0.42	3.03	0.41	2.48	0.33
	Overall	3.38	0.39	2.94	0.43	2.41	0.35

Table 3. Means and standard deviation of throughput for each device across both the right (R) and left (L) hands.

Performance Improvement

We performed t-tests to check for statistical significance between the performance in the first and second rounds measured in throughput. These tests revealed no statistically significant improvement in performance between rounds for the mouse on neither the right hand $(t(23) = 0.34, p = 0.63)$ nor the left hand $(t(23) = -0.94, p = 0.18)$. There were however improvements in performance between rounds with the touchpad for both the right hand, $(t(23) = -2.78, p < 0.01)$ and the left hand, $(t(23) = -5.41, p < 0.01)$. There was also an improvement with gestures for both the right hand, $(t(23) = -2.8542, p < 0.01)$, and the left hand $(t(23) = -3.81, p < 0.01)$.

To report the performance improvements between the rounds, a simple method would be to calculate the difference between the mean of means of throughput between both rounds:

$$\text{Improvement} = \frac{\text{Throughput}_{\text{R2}} - \text{Throughput}_{\text{R1}}}{\text{Throughput}_{\text{R1}}} \quad (8)$$

Hand	Mouse	Touchpad	Gestures
Right	-0.4%	4.4%	5.7%
Left	1.2%	6.1%	6.6%

Table 4. Performance increase measured as percentage of improvement in throughput between rounds

Improvements measured this way are listed in Table 4. This method while simple, was decided to not be a good reporting metric as it ignores relevant information such as standard deviation, which is important to be used in calculation to contextualize the difference [3]. We therefore decided that a better way to measure performance increase is with effect size. A standard metric used to report effect size is Cohen's *d* [4] which is defined as the difference of the mean between 2 groups over the standard deviation:

$$\text{EffectSize} = \frac{\overline{\text{x}_1} - \overline{\text{x}_2}}{\text{s}} \quad (9)$$

Effect Size measured as 0.2 has been considered a small effect, 0.5 signifies a medium effect visible to the naked eye, and 0.8 signifies a large effect size [3].

In the original equation above, the denominator is the standard deviation of the population, which would be impossible to obtain and was therefore replaced with the pooled variance of both samples. This is a common method used to calculate effect size and is used to assess performance improvement from one point to another [23, 14]:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (10)$$

The first thing to notice about the effect size listed in Table 5 is how the results for gestures can be interpreted differently based on the measures used. The simple performance improvement measured with percentage shows that left gestures has a higher improvement than right gestures, while the metrics calculated with effect size shows that right gestures is in fact higher.

Hand	Mouse	Touchpad	Gestures
Right	(0.05)	0.39	0.46
Left	0.11	0.41	0.45

Table 5. Performance increase as measured with Cohen's *d*. This measure generally gives a positive number in our experiment, but we noted that performance with the mouse on the right hand actually decreases.

A noteworthy observation is that the effect size differs very slightly between hands across all devices. For example the performance difference for the mouse is 0.05 and 0.11 for right and left hands respectively, touchpad is 0.39 and 0.41, while gestures' performance increase is measured as 0.46 and 0.45. This shows a very similar increase in performance using a particular device, regardless of the hand used to control the device.

We noticed that the performance improvement of the touchpad was closer to the gestures than the mouse despite the touchpad being a more ubiquitous interaction than gestures.

Degradation

Recall that we define degradation as the difference in performance between hands. To calculate degradation, we first used the original formula used by Kabbash et al [10]:

$$\text{Degradation} = \frac{\text{Time}_{\text{NPH}} - \text{Time}_{\text{PH}}}{\text{Time}_{\text{PH}}} \quad (11)$$

where TIME refers to means of movement time, whereas NPH and PH refers to non-preferred hand and preferred hand respectively. We found that this formula gives us degradation of 45% for the mouse, 27% for the touchpad, and 11% for gestures based on overall means of movement time.The full list of degradation for both rounds and over all 3 devices is available in Table 6.

Means	Mouse	Touchpad	Gestures
Round 1	45.2%	28.3%	11.9%
Round 2	44.6%	26.1%	10.5%
Overall	44.9%	27.3%	11.2%

Table 6. Degradation calculated as percentage of increase in movement time between hands

As movement time is a naive metric [30], we also performed the degradation calculation with mean throughput. A slight modification to the original formula in Equation 12 was required to allow for results to display positive numbers:

$$\text{Degradation} = \frac{\text{Throughput}_{\text{PH}} - \text{Throughput}_{\text{NPH}}}{\text{Throughput}_{\text{PH}}} \quad (12)$$

The results of this calculation shown in Table 7 provides a better metric, standardized with current literature, which recommends the use of throughput over movement time for benchmarking pointing devices [30, 31].

Source	Mouse	Touchpad	Gestures
Round 1	30.1%	20.1%	9.0%
Round 2	29.0%	19.0%	8.5%
Overall	29.6%	19.7%	8.7%

Table 7. Degradation calculated as percentage of decrease in throughput between hands.

In our calculation of performance improvement, we recommended the use of effect size and elaborated why it is a good choice in calculating differences between means of groups. We therefore used the same method to calculate degradation, listed in Table 8.

Source	Mouse	Touchpad	Gestures
Round 1	3.63	1.83	0.67
Round 2	3.44	1.60	0.71
Overall	3.57	1.69	0.68

Table 8. Degradation between hands measured with Cohen's *d*.

We observed that when using Cohen's *d*, degradation is reported as reduced between rounds for the mouse and the touchpad while it increases for gestures. Whereas the percent-wise calculation showed a decrease in degradation for all 3 devices as shown in Table 7. This is similar to the difference in our analysis of performance improvements, where reporting performance improvements with a percent-wise calculation gives different results than when Cohen's *d* is used.

Fatigue and Discomfort

Participants were asked in an exit survey whether or not they felt any fatigue or discomfort during the use of the gestural interface. 8 out of the 24 participant who used the gestural interface reported some degree of fatigue or discomfort.

ANALYSIS AND DISCUSSION

Literature Comparison

Although our task deviated slightly from the ISO 9241-9 standard for evaluating pointing devices our results were in line with previous literature using the standard. The reported throughput of 4.8 bits/s for the mouse using the preferred hand was consistent with [15] using the ISO 9241-9 standard for 2D multi-directional pointing evaluation which reported 4.9 bits/s. The throughput of the touchpad 3.66 bits/sec was much higher than the 2.9 bits/s previously reported [15]. It is possible that this is caused by hardware and software improvements made to the touchpad over the years, the difference in the dimensions of the device, as well as it's current ubiquity. Our gestural interaction had a reported throughput of 2.64 bits/sec, slightly better than previous reported gestural throughput of 2.10 bits/sec over an ID range 1-4 bits [18]. The degradation of the mouse calculated in percentage difference over movement time was 45%, much higher than the previous reported value of 27% [10]. This could be attributed to our participants using the computer and the mouse a lot more, as well as hardware and software improvements in the mouse since the study was conducted 21 years ago.

Our research showed a performance improvement between 2 rounds which contradicts a previous study [19]. This is likely due to the onset of fatigue in the aforementioned study, while we mitigate this issue in our research by allowing participants to rest their elbow on a surface.

Metrics

In this paper we focused on incorporating standard measures in reporting metrics, taken from peer reviewed literature outside and within the the realm of HCI. To start, we provided degradation between hands measured in throughput as opposed to movement time.

The impact of using Cohen's *d* as opposed to a percent-wise comparison is clearly illustrated in both performance improvements as well as degradation. Table 5 which reported performance improvement in Cohen's *d* showed higher improvement with gestures on the right hand than on the left hand, whereas Table 4 which reported performance improvement with percent-wise calculation showed higher improvement with the left hand. Likewise gestural degradation in Table 7 showed higher degradation in the first round when calculated in percentage but was lower in the first round when calculated with Cohen's *d*.

Cohen's *d* has been used to measure effect size in other areas of research including motor learning [14, 23, 21]. The use of this standard metric will be useful in comparing results across studies and has therefore been recommended to be used by HCI researchers [11]. Use of this metric would also facilitate replication and comparison of results in HCI, another aspect which has been recently encouraged [28].

Performance analysis of pointing devices have come a long way. We now know that it is best reported with throughput, and that movement time is a naive metric. Standardization in this area was favorable as it allowed researchers to compare results with that of other experiments. Likewise, we aim to allow future researchers to compare performance improvements between rounds, as well as degradation between hands. We propose reporting these values in effect size would be the best way to do just that.

Performance Improvement

It has to be expected that participants will be learning the pointing device evaluated as the experiment is being done. Our research showed that this was true of gestures and the touchpad, despite the participants being provided with a test run first. One way of reducing this effect is by making participants perform the same task repeatedly as done in [15]. In the aforementioned study, the participants were asked to perform the tasks over a period of time until no significant improvement in performance was noticed. Analysis of device performance was only performed on the subset of data where performance improvements were no longer noticed. This was not feasible in our experiment as each round required 4 tasks for the user to perform: 2 devices on both hands. Making participants perform the task until they have fully learnt the device could potentially take longer than the 1-2 hours recommended for usability studies [17]. To address this issue, we designed our experiment to cycle between tasks, thus allowing participants to gain long-term motor-learning benefits [14], and we measured these improvements with effect size.

We expect future researchers to face a similar issue. Simply performing one task per device is not indicative of the performance of that device. Providing a trial run may be beneficial, but as demonstrated by our experiments, it is still insufficient, as participants exhibited noticeable improvements despite the practice sessions. Performing a long term study until participants are fully familiar with the device may not be feasible due to time constraints and difficulty in retaining participants over the duration of the study. Performing at least 2 rounds on top of a practice session, and reporting performance improvements between the two rounds would help identify the extent to which the users are learning the device, as well as its future potential. For example when comparing 2 new interaction styles with similar performance, it would only be logical to pay more attention to one that exhibits higher performance improvements.

In our own experiments, we observed that performance improvement is more related to the interface than it is the hand. This tells us that interfaces themselves could have inherent performance improvement rates. We do not know if this is based on the participants' knowledge of the interface, or if it related to the design of the interface, however we do note that participants have better improvement on interfaces that are foreign to them (i.e. gestures) and lesser to no improvement on the devices that they are familiar with (i.e. mouse).

An interesting observation is the lack of performance improvement with the mouse, especially on the left hand. This has 2 potential explanations (1) that the mouse is an ideal design, allowing participants to use this device to it's full potential immediately or (2) that knowledge from the participants have acquired from long term use of the mouse on their preferred hand has transferred to their non-preferred hand. The second hypothesis is more likely as it has been shown in other studies that motor skill learned on one hand is transferable to the other [7].

The fact that gestures had a higher performance improvement on both hands compared to the other devices validates the gestural interaction style [9] used here. It also indicates that there is potential for the use of gestures as a pointing devices, despite the lower throughput recorded, as participants were still learning the interaction. As performance improvements were not noticed on the mouse, we can infer that any improvements recorded are the result of learning the interaction and not the task. We therefore conclude that the higher performance improvement with gestures reflect the potential of this interaction. Future study of this interaction style will be required to find out the number of sessions required before performance improvements are no longer noticed.

Degradation

Devices that exhibit high degradation will behave noticeably different on the preferred hand than on the non-preferred hand. This difference could discourage the users from using the device on their non-preferred hand as they already have an expectation in performance. Since gestures had a much lower degradation than the other devices, we expect it to perform more similarly on the non-preferred hand, thus allowing for 2-handed interactions.

We have identified two potential explanations for the degradation in different devices, (1) touchless gestural interaction is natural enough to be used with the non-preferred hand or (2) participants' pre-existing motor skills with regards to the device caused a higher degradation on interactions they were more familiar with. A concrete conclusion is hard to make as there is little known on the degradation with other interfaces used in HCI.

Performance degradation between hands is a useful metric to report when assessing pointing devices because it allows researchers to identify interfaces, interactions or tasks that may be suitable for the non-preferred hand. We report degradation in effect size as it allows for a much more meaningful interpretation; simply reporting the difference in percentage ignores the standard deviation and could result in different results, as illustrated by the difference between Tables 7 and 8.

Gestures

Our gestural interaction design was based on the Personal Space approach [9]. This study showed that allowing the users to rest their elbow on a surface would severely reduce fatigue while maintaining performance to a similar level as in a standard gestural interaction approach where participants held their hand out. It was reported that the standard approach resulted in fatigue after approximately 90 seconds whereas there was no fatigue reported in the Personal Space approach.

A previous study in assessing gestural interaction showed no improvement in performance between rounds, a direct contradiction of our findings [19]. This previous study also reported that fatigue did not have any effect, despite the use of the standard approach to gestural interaction. There are a few explanations why these researchers did not observe any performance improvements in their experiments: (1) The interaction style used in our research which allowed the participants to rest their elbow allows for a better motor skills improvement. (2) The performance increase facilitated by increase in motor skills is negated by fatigue, leading to zero net gain. We believe the second explanation is more likely given that it has been demonstrated that the standard interaction style would cause fatigue. This experimental design had 5 blocks of 20 target selection sequences, with a mandatory 10-second breaks. This design may have caused for there to be no fatigue reported, despite its presence.

CONCLUSION

In this paper we examined gestural interaction for throughput, performance improvement and degradation, benchmarked against the mouse and touchpad. Although our study shows that gestures performs lower than the mouse and the touchpad, it does have much higher performance improvement between 2 rounds and much lower degradation between hands. The high performance improvement indicates that gestural interaction has a good potential for use in productive use such as on the desktop, as participants will learn the interaction over time. While low degradation indicates that it is possible to use gestures in 2-handed interactions or to allow users to easily alternate between hands when used as a pointing device.

REFERENCES

1. Bødker, S. When second wave hci meets third wave challenges. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*, NordiCHI '06, ACM (New York, NY, USA, 2006), 1–8.
2. Carmody, T. Why 'gorilla arm syndrome' rules out multitouch notebook displays. *Wired, Oct 10* (2010).
3. Coe, R. It's the effect size, stupid: What effect size is and why it is important. *Paper presented at the British Educational Research Association annual conference, Exeter, UK* (2002).
4. Cohen, J. A power primer. *Psychological bulletin 112*, 1 (1992), 155.
5. Fitts, P. M. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology 47*, 6 (1954), 381.

6. Goble, J. C., Hinckley, K., Pausch, R., Snell, J. W., and Kassell, N. F. Two-handed spatial interface tools for neurosurgical planning. *IEEE Computer 28*, 7 (1995), 20–26.

7. Grafton, S. T., Hazeltine, E., and Ivry, R. B. Motor sequence learning with the nondominant left hand. *Experimental Brain Research 146*, 3 (2002), 369–378.

8. Jacob, R. J., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T., and Zigelbaum, J. Reality-based interaction: a framework for post-wimp interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2008), 201–210.

9. Jude, A., Poor, G. M., and Guinness, D. Personal space: User defined gesture space for gui interaction. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, ACM (New York, NY, USA, 2014).

10. Kabbash, P., MacKenzie, I. S., and Buxton, W. Human performance using computer input devices in the preferred and non-preferred hands. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, ACM (1993), 474–481.

11. Kaptein, M., and Robertson, J. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 1105–1114.

12. Keir, P. J., Bach, J. M., and Rempel, D. Effects of computer mouse design and task on carpal tunnel pressure. *Ergonomics 42*, 10 (1999), 1350–1360.

13. Lee, T. D., and Genovese, E. D. Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport 59*, 4 (1988), 277–287.

14. Lin, C.-H. J., Sullivan, K. J., Wu, A. D., Kantak, S., and Winstein, C. J. Effect of task practice order on motor skill learning in adults with parkinson disease: a pilot study. *Physical therapy 87*, 9 (2007), 1120–1131.

15. MacKenzie, I. S., Kauppinen, T., and Silfverberg, M. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2001), 9–16.

16. Norman, D. A., and Nielsen, J. Gestural interfaces: a step backward in usability. *Interactions 17*, 5 (2010), 46–49.

17. Olsen Jr, D. R. Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, ACM (2007), 251–258.

18. Pino, A., Tzemis, E., Ioannou, N., and Kouroupetroglou, G. Using kinect for 2d and 3d pointing tasks: performance evaluation. In *Human-Computer Interaction. Interaction Modalities and Techniques*. Springer, 2013, 358–367.

19. Sambrooks, L., and Wilkinson, B. Comparison of gestural, touch, and mouse interaction with fitts' law. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, ACM (2013), 119–122.

20. Schmidt, R. A., and Lee, T. *Motor Control and Learning, 5E*. Human kinetics, 1988.

21. Seidler, R. D. Multiple motor learning experiences enhance motor adaptability. *Journal of cognitive neuroscience 16*, 1 (2004), 65–73.

22. Soukoreff, R. W., and MacKenzie, I. S. Towards a standard for pointing device evaluation, perspectives on 27 years of fitts law research in hci. *International Journal of Human-Computer Studies 61*, 6 (2004), 751–789.

23. Sullivan, K. J., Knowlton, B. J., and Dobkin, B. H. Step training with body weight support: effect of treadmill speed and practice paradigms on poststroke locomotor recovery. *Archives of physical medicine and rehabilitation 83*, 5 (2002), 683–691.

24. Teixeira, V. *Improving elderly access to audiovisual and social media, using a multimodal human-computer interface*. PhD thesis, Faculdade de Engenharia, Universidade do Porto, 2011.

25. Wachs, J. P., Kölsch, M., Stern, H., and Edan, Y. Vision-based hand-gesture applications. *Commun. ACM 54*, 2 (Feb. 2011), 60–71.

26. Wachs, J. P., Kölsch, M., Stern, H., and Edan, Y. Vision-based hand-gesture applications. *Communications of the ACM 54*, 2 (2011), 60–71.

27. Wigdor, D., and Wixon, D. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier, 2011.

28. Wilson, M. L., Mackay, W., Chi, E., Bernstein, M., Russell, D., and Thimbleby, H. Replichi - chi should be replicating and validating results more: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, ACM (New York, NY, USA, 2011), 463–466.

29. Wu, Y. What else should an hci pattern language include. *Procs. of Pattern Languages for Interaction Design: Building Momentum@ CHI* (2000).

30. Zhai, S. Characterizing computer input with fitts law parametersthe information and non-information aspects of pointing. *International Journal of Human-Computer Studies 61*, 6 (2004), 791–809.

31. Zhang, X., and MacKenzie, I. S. Evaluating eye tracking with iso 9241-part 9. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*. Springer, 2007, 779–788.