Learning and Performance with Gesture Guides

Fraser Anderson
University of Alberta
Edmonton, Alberta, Canada
frasera@ualberta.ca

Walter F. Bischof
University of Alberta
Edmonton, Alberta, Canada
wfb@ualberta.ca

ABSTRACT
Gesture-based interfaces are becoming more prevalent and complex, requiring non-trivial learning of gesture sets. Many methods for learning gestures have been proposed, but they are often evaluated with short-term recall tests that measure user performance, rather than learning. We evaluated four types of gesture guides using a retention and transfer paradigm common in motor learning experiments and found results different from those typically reported with recall tests. The results indicate that many guide systems with higher levels of guidance exhibit high performance benefits while the guide is being used, but are ultimately detrimental to user learning. We propose an adaptive guide that does not suffer from these drawbacks, and that enables a smooth transition from novice to expert. The results contrasting learning and performance can be explained by the *guidance hypothesis*. They have important implications for the design and evaluation of future gesture learning systems.

Author Keywords
Gestures, guides, evaluation, learning

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces. - Theory and methods; Training, help and documentation; Input devices and strategies.

General Terms
Design**;** Human Factors

INTRODUCTION

Gestural interfaces are becoming widespread as the adoption of interactive surfaces, touch screens, and tablets increases. The use of on-screen gestures allows displays to be multiplexed as both input and output devices. Gestural input is also able to leverage the rich capabilities of the human motor system, allowing for efficient interaction. The difficulty with gesture-based interfaces has been the learnability of the systems, as gestures are inherently difficult to discover and predict without an explicit guide [19, 39]. Learnability is thus a critical issue, as efficient interaction relies on the ability to execute a large set of memorable gestures.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2013, April 27 – May 2, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1899-0/13/04...$15.00.

Learnability of gesture sets involves two factors. The first is the cognitive mapping between the desired action and the required gesture. This paired-associate type of learning is what is typically studied in gesture learning. The second, equally important aspect of gestural interactions is the ability to perform a gesture. Bau and MacKay recognize the importance of the gesture execution, stating that users must "*master the details of drawing the shape to improve recognizer accuracy*" [4]. This component of gestural interaction becomes increasingly important as the use of gestural interfaces continues to grow, and devices rely solely on gestural input. In the case of experts, many of their input sequences are largely automatic, relying primarily on responses from the motor system. Motor performance is important for novices as well. As the size of gesture sets is increasing (e.g, 40 targets [23]), both novices and experts have to perform gestures with increasing accuracy in order for the recognizer to distinguish them from other, potentially similar gestures. It is also foreseeable that advanced gestural interfaces will allow users to modify parameters of commands by producing variations on gestures, which again require some skill to perform.

Researchers have acknowledged the difficulty in using complex gestural interaction systems and have developed a number of systems designed to improve the usability and learnability of gestural interfaces. As early as 1994, Kurtenbach, Moran and Buxton developed animated crib notes to help users learn to perform gestures [19]. While crib notes alone would be sufficient to aid users in the recall of gestures, the addition of in-context animations provided extra cues that helped users in the execution of the gesture. Recently, systems offering dynamic, real-time guidance have been proposed [2, 5, 4, 13, 16]. These systems provide the user with information to help guide the execution of a gesture, and they have been seen as an improvement to traditional help menus or gesture demonstrations. These guides are believed to improve performance, as "*feedforward and feedback facilitates learning and execution of complex gesture sets*" [4].

To date, most studies have evaluated the learnability of gestural interfaces by comparing performance measures taken during, or shortly after, the training phase of an experiment [3, 4, 7]. These often include gesture recall, frequency of gesture use, or input speed. Although this evaluation is direct and intuitive, as it mimics real world use cases, it does not necessarily evaluate how well the participants learned the gestures.

The field of motor learning has established methods for assessing the ability to learn and execute movements. This

field is concerned with a wide range of activities, from simple movements (e.g., pointing and grasping [10]) to complex skills such as surgical movements [8] and sports [15]. The motor learning literature acknowledges a critical difference between performance and learning [31]: Performance refers to the *production of a specific action*, whereas learning refers to the *relatively permanent acquired capabilities that facilitate improved performance*. Empirical studies that separate performance from learning involve a training phase followed by a retention component and a transfer component. In the retention component, participants perform the task at a common level of the independent variable, typically 24 to 48 hours after training. In the transfer component, participants perform a novel variation on the task they were trained on, e.g., performing the task with the other limb.

The separation between learning and performance becomes even more important when considering the guidance hypothesis [30]. The guidance hypothesis states that excessive guidance during training can hinder learning, as the user can become reliant on the guidance. Guidance can take the form of knowledge of results (KR), which is information regarding the success or failure of a movement, or knowledge of performance (KP), which is information regarding how the participant and target movements differ. The amount of guidance provided to a user is an important consideration, as many new gesture guides provide concurrent or real-time feedback (or 'feedforward') to help the user execute the gesture.

This paper makes three main contributions to the literature on gesture learning. First, we introduce the use of the retention and transfer experimental paradigm within the context of gesture learning. Second, we analyze four different gestural guides with this paradigm. We find that guides with high performance during training result in poor performance during retention and transfer, and the converse for guides with low performance during training. This indicates a tradeoff between ease of use and learning. Third, we introduce an adaptive guide that mitigates this tradeoff, and provides a smooth transition from novice to learned user. This work has important consequences for the design and evaluation of future gestural interfaces.

RELATED WORK

Evaluating Gestural Interfaces

One approach to evaluate gesture-based systems is to analyze behavior while participants are using the gesture system. In these studies, researchers typically analyze the frequency with which the gestures are used, the rate of gesture input, or user preference with the gesture system [6, 21]. Appert and Zhai analyzed preference and memorability for keyboard shortcuts and gestures after training with both systems [3]. They found users did not have to consult the help menu system as often with gestures, and the use of gestures resulted in faster and more accurate recall of menu commands. To evaluate their menu-based gestural learning system, GestureBar, Bragdon et al. analyzed the number of correct gestures and the number of attempted gestures as participants used a gestural diagram editor [7]. Kurtenbach et al. evaluated the performance improvements over time as participants learned to use marking menus [18]. These systems evaluate user behavior while they are actively using the system, but do not separate performance from learning.

Another approach to evaluate gesture systems is to evaluate the ability to recall specific gestures after training [6]. To evaluate their dynamic and traceable gestural guide, OctoPocus, Bau and MacKay compared participants' ability to recall gestures before and after training with a gesture system and a traditional help-window [4]. In the evaluation of a multi-touch gestural guide system, ShadowGuides, participants recalled gestures immediately following a training phase with ShadowGuides or a video-based guide [13]. Zhai and Kristensson extended this paradigm over several days of training and testing to evaluate the ability of participants to recall 100 gestures [42]. Though the stated aim of some of these systems is to assist users in the performance or execution of the gesture, they tend to focus on the cognitive component of gesture learning as measured by recall. While recall is a useful measure to assess the degree to which the action-gesture pairing was learned, these studies do not analyze the motor component of the gesture and are not able to fully isolate performance from learning.

A third approach to evaluate gesture systems is to achieve peak performance on a subset of representative gestures, and use this to predict expert behavior [17, 43]. This approach evaluates the potential bandwidth of a gestural interaction system, but does not assess gesture learnability.

Retention and Transfer

The use of retention and transfer tests is standard in motor learning literature, as they allow the researcher to separate the effects of performance factors from learning factors [32]. Performance factors have effect only for a short time, whereas learning factors have effect much longer after training. While it is commonly believed that increased performance on practice trials is indicative of learning, Schmidt and Bjork [32] point out that this is not the case. Drawing from examples of motor and verbal tasks, they show that performance during practice is not indicative of learning, but only evident through retention and transfer tests.

Typically, researchers use one or more retention tests, with a delay of 15 minutes, 1 hour, or more than a day later [35]. Tests are frequently performed after at least one full night of sleep (e.g., 24 hours), as sleep has been shown to play an important role in the consolidation of motor skills [29]. The task performed during retention tests is usually similar to the task performed during training, but with all participants moved to the same level of the independent variable, which is often the removal of all feedback [31].

Transfer tests are another way to assess learning, as the mental changes associated with learning one skill will frequently be generalizable to another, very similar skill. These tests can also show how well a learned skill generalizes to a new context. In a transfer task, participants are asked to

perform the trained task, but with a novel variation. For instance, participants might be asked to perform a learned skill at a different scale or angle [1]. These types of tests are widely used in motor learning, and many skills show substantial transfer.

There has also been substantial work within the motor learning field on bilateral transfer, i.e., the degree to which a skill learned on one hand transfers to the other. While there are no steadfast rules on what parameters or types of skills transfer from one hand to the other, it is clear that there is often a substantial amount of transfer that takes place [25, 28]. These studies show that transfer takes place even when participants are not ambidextrous. While bilateral transfer is not of direct and obvious benefit to many HCI scenarios, it is perhaps the simplest and most direct manipulation to assess transfer. It requires no software modifications, reducing potential errors in the implementation of the study. Additionally, the purpose of the transfer task is not to assess the applicability of the gesture to other contexts, but simply to assess the degree to which the gesture was learned.

Guidance and Effort

Studies of the guidance hypothesis have explored the effects of various types of feedback on performance and learning. Early studies on guidance focused on the guiding effects of KR [38]. Excessive KR encourages participants to perform "maladaptive short-term corrections" [30] rather than to learn more effective corrections that are useful in the longer term. For example, participants showed greater learning when KR was presented after a series of trials, compared to being presented after each successive trial (terminal feedback) [34]. More recently, the guiding effects of KP have been shown to be stronger than those of KR [20]. Park, Shea, and Wright showed that concurrent feedback alone provides a much stronger guidance effect than terminal feedback, with higher performance during training and much lower performance during retention [24]. Schmidt and Wulf found similar results when analyzing the ability to learn a spatiotemporal movement pattern with or without concurrent feedback [33]. These studies have important implications for gesture learning as many gesture guides present concurrent KP in the form of real-time assistance as the gestures are being performed.

In a study of gesture-based text entry [11], researchers manipulated the amount of effort required to view the keys on a gesture-based keyboard. They found that the interface requiring more effort substantially slowed down training and resulted in greater accuracy in remembering the spatial location of targets than the less-effortful interface. Similar effects have been found during cognitive tasks. van Nimwegen and Oostendorp found guiding effects when providing automated assistance in solving a constraint satisfaction task [36]. Rappin et al. found students learned more in an interface that forced them to interact with a chemical simulation, rather than simply observe it [26]. These and other studies on guidance in human computer interfaces focus on increased learning due to effort or cost of interaction [11, 14, 26, 36]. In contrast, the results from our experiments using a retention and transfer paradigm show an effect which is not based on effort, and can be better explained by the guidance hypothesis.

In the present work, we unify the existing work in motor learning and human-computer interaction in order to better study gestural interfaces. By applying the standard evaluation paradigms from motor learning, we are able to study how the guidance hypothesis affects gesture learning and execution.

METHODS

Participants

A total of 36 subjects participated in the study (15 male, 18-77 years, M=25 years, SD=11 years). While research with older users presents its own challenges [41], only one participant was over 60 years of age, and their data was not abnormal. All participants were right-handed, as determined by the Edinburgh handedness inventory [22]. Each participant was assigned to one of four gesture guides: *cribnotes*, *static-tracing*, *dynamic-tracing*, or *adaptive* which are described below. Before beginning the training phase, each participant was informed that there would be a follow-up test, but were not informed of the nature of this test.

Apparatus and Gestures

The experiment was performed using a pen-based Cintiq 21UX from Wacom, with the screen positioned directly in front of the participant at a 10° incline. The software was developed using the Windows Presentation Framework, and ran full-screen at 1600 x 1200 pixels. For reference, the display size of the screen is (43 cm x 32.5 cm), resulting in a mapping of 1 px = 0.27 mm. The buttons on the pen were not used; all interaction was accomplished through the contact and motion of the pen on the screen.

Each of the four gestures (see Figure 1b) was composed of one or two simple line or curve segments and paired with an arbitrary, unrelated verb. All gestures were the same length and defined by a single, computer-generated template rather than as a series of user-generated examples. The gestures were designed to cover a wide range of possible gestures, as all gestures can be described as a series of curves, lines, and corners [9]. The initial angle of each gesture was rotated such that it did not coincide with a major axis or a diagonal. While typical gesture systems use more than four gestures, this study is intentionally restricted to analyzing only four. If more gestures are added, participants have a difficult time learning the pairing between gesture and command, i.e., they struggle at the 'cognitive' phase of Fitts and Posner's model rather than progressing through to the associative or autonomous phases [12]. As the intent is to study the production and form of the gesture, the use of four gestures allows participants to quickly learn the pairing so they can then better learn the form.

The red gesture, paired with '*Choose*', was two straight lines of equal length joined by an obtuse angle. The green gesture, paired with '*Send*', was a curve of constant radius, sweeping out a 180° arc. The blue gesture, paired with '*Build*', was a curve of constant radius connected to a

straight line. Lastly, the orange gesture, paired with '*Find*', was composed of a long straight line connected to a short straight line at a 90° angle.

Guide types

We evaluated four types of gesture learning systems. Three of the guides have been previously described in the literature or are very similar to previously described guides (*crib-notes*, *static-tracing*, and *dynamic-tracing*), while the fourth (*adaptive*) is a novel contribution.

The *crib-notes* guide used a half-scale depiction of the gestures placed in the top-left corner of the screen (Figure 1a). Participants using this system were not informed of the scale relation between the guide and the target gesture, and learned the appropriate scale through the KP provided after each trial. This guide provides the least guidance, as participants cannot directly compare their current trajectory to the template and the template does not adapt to their movement.

The *static-tracing* guide used a full-scale depiction of each of the template gestures, radiating from the initial pen location (Figure 1b). This guide allowed the participant to trace over the target gesture. As the participant drew their stroke, the guide was not updated in any way. The use of this guide allowed us to examine what effects the continuous updating has on the learning and performance of the gestures.

The *dynamic-tracing* guide (referred to as 'dynamic guide' in Bau and MacKay [9]) used a full-scale depiction of the gestures, as with the static-tracing guide, but as the participant moved the pen, the guide dynamically updated to reflect the state of the recognizer (Figure 1c). As the participant drew their stroke, the opacity of each of the four gestures was mapped to a function of the similarity between the participant's trajectory and the template of the target gesture. Gesture similarity during training was measured by computing the RMSE between the participant's trajectory and an equivalent path length from each of the target gestures. In addition to modifying the opacity of the guide strokes, the initial segment of each the template gesture was removed (an amount equal to the current participant's stroke length), and the result is appended at the current pen location. This procedure effectively provided the '*feedforward*' information to help guide the participant to the correct performance.

The *adaptive* guide provided a traceable guide identical to the one used in the static-tracing condition, but the guide disappeared at some point in time during the trial. The current trial as well as the current length of the participant's stroke determined when the guide disappeared. For the first trial, the guide disappeared once the participant's stroke had the same path length as the target gestures. Midway through the trials, the guide disappeared once the participant's stroke was half the path length of the target gestures. By the end trial, the gesture guide did not appear at all. This approach let participants initially trace the gestures with high accuracy and usability, but eventually required them to draw the gestures without the guide. While the implementation of this guide for the lab study is straightforward, as the number of trials is known, the implementation in a real-world scenario is potentially more difficult. Various methods of implementing an adaptive guide in a real-world scenario are described in the Discussion section.

All of the guides used in this study were not dynamic in the sense that they changed scale or orientation in response to the user's strokes, as in other recent guide designs [2,25]. This is an intentional choice, as it allows control over the exact gesture being learned by the participants. This decision allows more precision in studying the effects of the guide on learning a particular gesture. It is highly unlikely that the ability to change scale or orientation will have any effect on the degree to which the user is guided, or subsequently learns the gesture. Once a user determines or selects a particular scale and orientation, they will likely use the guide to continue drawing the gesture at that particular scale or orientation. That is, the users would still be guided to the same degree, they would just be guided to a different target gesture.

Procedure

Participants were shown where to place the pen on the screen to activate the guide and where their score would appear. They were told to accrue as many points as possible and that their score was derived from the similarity to the target gesture, with additional points for faster performance. To compute the points, the training system awarded

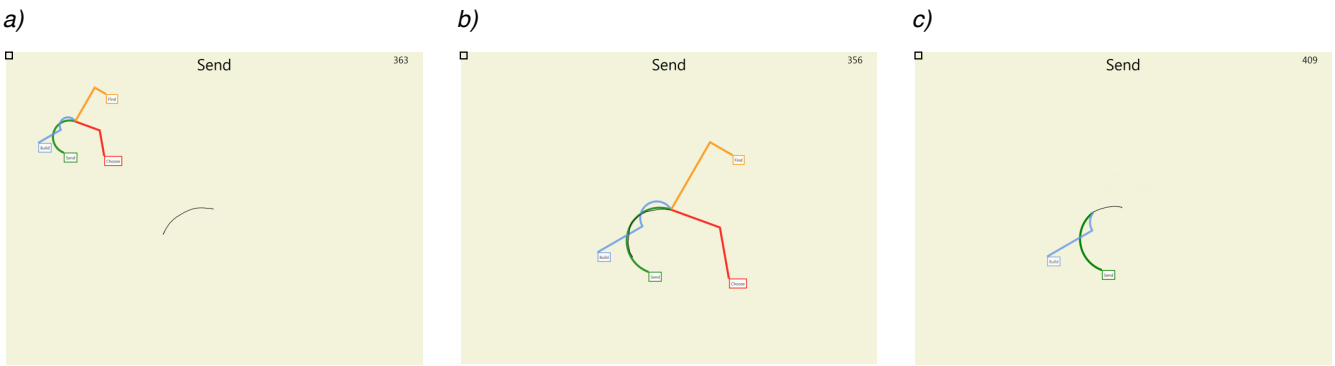

Figure 1: Behavior of the guides while performing the *Send* gesture during training trials for a) crib-notes, b) static-tracing, and c) dynamic-tracing. Note that adaptive is not shown as its behavior is identical to the 'static-tracing' guide, except the guide is removed partway through the trial.

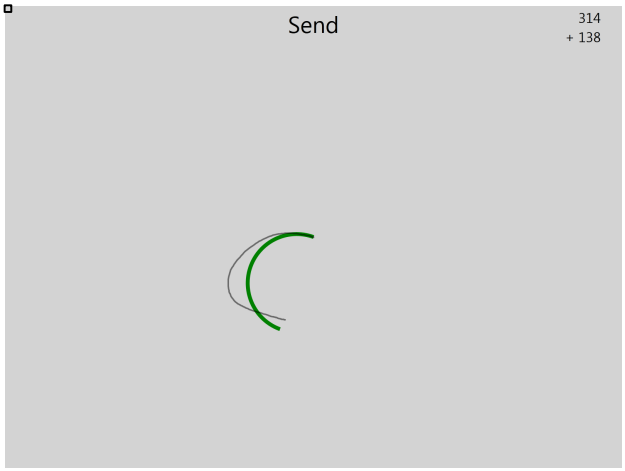

Figure 2: Inter-trial screen showing KR (score in top right) and KP (trajectory overlay).

the user with points proportional to an execution time under four seconds and an accuracy error under 220 pixels. For instance, if the gesture was completed in two seconds with 30 pixels of error (i.e., average crib notes performance halfway through the trials) the participant received 197 points.

The training phase consisted of 60 trials for each gesture, with the presentation order randomized such that no gesture appeared in more than two successive trials. At the start of each trial, the target word appeared at the top of the screen. The current score, as well as the target circle, were also visible. To begin the trial, the participant placed the pen tip in the target circle, which caused the gesture guide to immediately become visible. The participant then drew the gesture on the screen, which left a visible 'ink trail'. When the pen was lifted, all on-screen content was hidden for 1000 ms. The participant was then provided with KP and KR, consisting of the target gesture along with the participant's trajectory overlaid and the score for the current stroke (Figure 2). After 1500 ms of exposure to the KP and KR, the screen went blank for 1000 ms, and then the next target word appeared. All training was performed with the right hand.

Participants performed retention and transfer tests 15 minutes after completing the training phase. The retention test was similar to the training, with each participant performing 16 trials (4 per gesture, with the order randomized), but no guide (KR or KP) was provided. Participants were instructed to draw the gestures from memory. They were also reminded of the four target words and told to take as much time as needed before drawing the gestures. The participants were not shown any of the gestures. The transfer test was identical to the retention test (16 trials with no guide, KP, or KR), but performed with the left hand. Approximately 24 hours later, each subject completed the retention and transfer follow-up tests again.

RESULTS

Training, retention, and transfer data was analyzed with three mixed-design ANOVAs, and post-hoc tests were conducted using Tukey's HSD. Prior to each analysis, trials where the participant performed the incorrect gesture were discarded, as the focus of this study was on the form of the gesture, not on gesture recall. In other words, these errors were removed as we were interested in errors due to performance (i.e., slips) rather than cognitive errors (i.e., mistakes). This resulted in less than 1% of data being removed from training, retention, and transfer phases. These errors were spread evenly across all guides, and came primarily from the training data.

Gesture Similarity

The similarity of each stroke to the template was computed by resampling the template and the participant strokes to 128 evenly spaced points, then computing the root mean square error (RMSE) of the Euclidean distance between corresponding points. This method is sensitive to both scale and rotation, as participants were instructed to match the template gestures along those dimensions as well as shape. A graphical representation of the similarity data separated by GuideType is shown in Figure 3, and by gesture is shown in Figure 4.

While RMSE is not the most popular method in gesture recognition, there are several reasons that make it a good choice for analyzing accuracy of motor production. First, the use of RMSE gives a direct and accurate measurement of the participant's ability to produce the target gesture. Secondly, it does not rely on a collection of high-level features (see, e.g., Rubine's algorithm [27]), the selection of which will change with the next advancement in gesture recognition. That is, our results are independent of the current state of the art in gesture recognition. It is also worth noting that the error was also analyzed using the error measure used for the $1 recognizer [40], as well as by using the number of 'points' awarded each trial, and the same patterns emerged from the resulting data.

Training

The training data was blocked by averaging six consecutive trials for each gesture, resulting in a total of 10 training blocks. As the error distributions were skewed, a log transform was applied to the RMSE values to normalize them before conducting the ANOVA. The same log transform was applied to all RMSE values before the analysis in subsequent sections. A 3 (GuideType) x 4 (Gesture) x 10 (Block) mixed-design ANOVA was conducted with Gesture and Block as within-subjects factors and GuideType as between-subjects factor (summarized in Table 1).

Post-hoc tests on GuideType showed that all guide types produced significantly different scores during training. From lowest to highest error produced, the guides are: 'Static', 'Dynamic', 'Adaptive', 'Crib'. Participants generally improved during training, with the error in the first block being significantly higher than the last block. The exception to this is with the adaptive guide, where participants progressively decreased in performance, due to the guide being removed earlier in the trial as they progressed through the training.

The main effect of gesture shows that the '*Find*' gesture was significantly easier to perform than the '*Choose*' and

'*Send*' gestures, and *'Build'* was easier to perform than *'Send'*, but the effect sizes were very small, and therefore were not analyzed further.

Retention

To analyze the retention data, the four trials for each gesture were averaged per participant and analyzed using a 4 (GuideType) x 4 (Gesture) x 2 (Delay) mixed-design ANOVA with GuideType as between-subjects factor and Gesture and Delay as within-subjects factors. Gesture was not found to be significant ($F_{3, 96} = 0.30$, $p = 0.83$), so this factor was pooled and the ANOVA was re-computed.

Both Delay ($F_{1, 248} = 8.90$, $p = 0.0031$, $\omega^2 = 0.02$) and GuideType ($F_{3, 32} = 2.93$, $p = 0.049$, $\omega^2 = 0.09$) were significant. Participants trained with crib-notes or the adaptive guide had significantly lower retention scores than those trained with either of the traceable guides. There was no significant difference between the retention scores of participants trained with either of the traceable guides. There was also no significant difference in the retention scores of participants trained with the adaptive guide or the crib notes. Performance on the 24 hour follow-up was poorer across all participants, compared to the 15 minute follow-up.

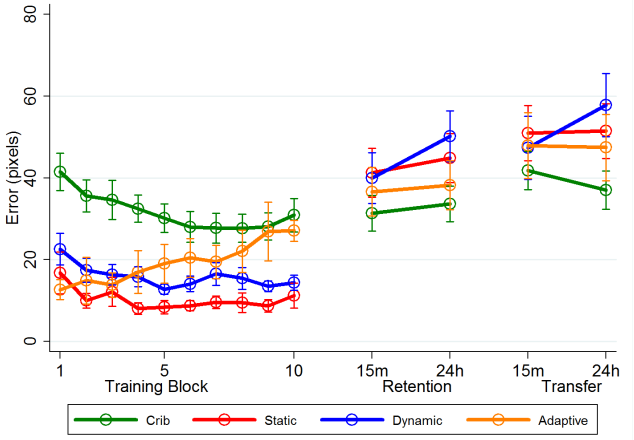

Figure 3: Error for training, retention and transfer, for each guide type. There is an apparent tradeoff between performance during training, and the amount of learning, measured by retention and transfer.

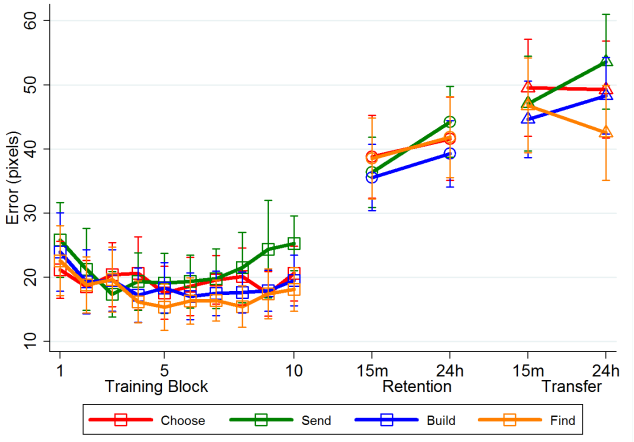

Figure 4: Error for training, retention, and transfer, for each gesture. There are no appreciable differences between the performance of each gesture. Error is primarily due to the type of training guide used.

Factor	*F*	*p*	ω^2
GuideType	$F_{3,32} = 34.34$	0.00	0.45
Gesture	$F_{3,96} = 4.20$	0.01	0.01
GuideType x Gesture	$F_{9,96} = 1.44$	0.18	0.01
Block	$F_{9, 288} = 5.08$	0.00	0.05
Block x GuideType	$F_{27,288} = 7.45$	0.00	0.06
Block x Gesture	$F_{27,864} = 0.15$	0.12	0.00
GuideType x Block x Gesture	$F_{81,864} = 1.08$	0.31	0.00

Table 1: ANOVA results for the training similarity data.

Transfer

To analyze the transfer data, all four trials for each gesture were averaged and analyzed using a 4 (GuideType) x 4 (Gesture) x 2 (Delay) mixed-design ANOVA with GuideType as between-subjects factor and Gesture and Delay as within-subjects factors. Again, the Gesture factor was not significant ($F_{3, 96} = 1.23$, $p = 0.30$) and pooled in the reported results.

The transfer results mimic the same pattern as the retention results, as evidenced by a Pearson's correlation ($\rho = 0.85$, $p < 0.001$). While the ANOVA did not report significant main effects, this similarity in results to the retention results demonstrates the potential utility of transfer scores. One reason for the lack of significant main effects is the performance improvement in the crib-notes trained participants following the 24 hour rest period, contrasted with the decreased performance of the participants trained with the dynamic guide.

Duration

Duration was measured as the time from the pen's initial contact with the screen to when the pen left the screen. This measure includes any time the participant spent consulting the guide as well as the time to draw the gesture. Duration data for the training, retention, and transfer phases are shown with results separated by GuideType in Figure 5, and results separated by Gesture in Figure 6.

Training

The mixed design ANOVA showed a main effect of Block ($F_{9, 216} = 25.5$, $p < 0.001$, $\omega^2 = 0.13$). There was a significant decrease in duration with nearly every block. There was also a main effect of Gesture ($F_{3, 72} = 19.3$, $p < 0.001$, $\omega^2 = 0.01$). The 'Send' and 'Find' gestures were both performed significantly faster than the 'Choose' and 'Build'

gestures, but the effect size is very small. GuideType had no effect on duration ($F_{2, 24} = 0.53$, $p = 0.59$, $\omega^2 = 0.02$).

Retention

The retention data shows only a main effect of Gesture ($F_{3, 96} = 3.95$, $p = 0.01$, $\omega^2 = 0.02$). 'Choose' is performed significantly slower than 'Send' and 'Find', but again, the effect size is quite small. As with the training data, the guide type has no effect on duration ($F_{3, 32} = 0.60$, $p = 0.62$, $\omega^2 = 0.03$).

Transfer

There was a main effect of Gesture during transfer ($F_{3, 96} = 15.86$, $p < 0.001$, $\omega^2 = 0.04$). Post hoc tests revealed that all gestures were significantly different from each other. In increasing order of duration, the gestures are: 'Send', 'Find', 'Choose', and 'Build'. These results were very similar to the training durations, except that the variance between gestures is increased, resulting in higher significance.

DISCUSSION

The type of guide used during training has clear effects on the behavior during training, as well as the retention and transfer scores. With high performance during training, there is little learned, but with low performance the participants retained substantially more. Additionally, by adapting the guide over time, the participants are able to balance performance and learning. These results have important implications for both the design of gestural guides, as well as for the way they are evaluated.

Guide Design

The three 'traditional' gesture guides (i.e., *static-tracing, dynamic-tracing*, and *crib notes*) showed performance improvements with training. There was an obvious and significant difference during training between crib-notes and the traceable guides with the crib-notes-trained participants performing much worse. Looking solely at the training data, it appears that the traditional guide types provided equal amounts of learning, but the baseline performance for crib-notes was worse. However, this was not the case. The retention scores, which mimic an expert-usage scenario showed the lowest error for crib-notes.

The newly proposed adaptive guide shows a much different result. While the traditional gesture guides lead to improved performance during training, the adaptive guide shows a gradual decrease in performance. This performance loss is easily explained. As the participant completed more trials, the guide disappeared earlier and earlier during gesture execution, forcing them to perform more of the gesture without the guide in place. In contrast to the traceable guides (static and dynamic), the adaptive guide actually provides a relatively smooth transition from novice to expert; there is not a substantial decrease in performance when the guide is removed. In contrast to the crib-notes guide, the adaptive guide provides a much more usable interface to novices, allowing direct tracing and high accuracy at the beginning of the training phase.

With respect to the three traditional guides, it seems that the more guidance given during training, the worse the learning. The participants using traceable guides had the most guidance and the worst performance in retention and transfer. Conversely, the participants using crib-notes had the least guidance during training but the best performance during retention and transfer. These results are explainable by the guidance hypothesis, and only become apparent within the retention and transfer paradigm. This shows an apparent tradeoff between immediate performance and learning. In addition, it appears that the dynamic guide has no benefit for performance or learning. The dynamic-traceable guide resulted in worse performance during training than the static guide, and users of this guide showed little to no learning of the gestures during retention and transfer. This is interesting, but not entirely surprising. That is, anytime the dynamic guide updated, it would necessarily deviate from the template trajectory. Thus, participants who

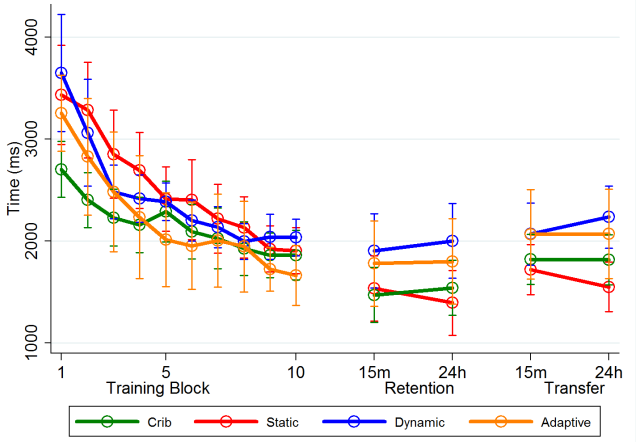

Figure 5: Duration of stroke during training, retention, and transfer, for each guide type. Guide type has little effect on the speed, allowing the tracing-based guides to provide high accuracy without a decrease in input speed.

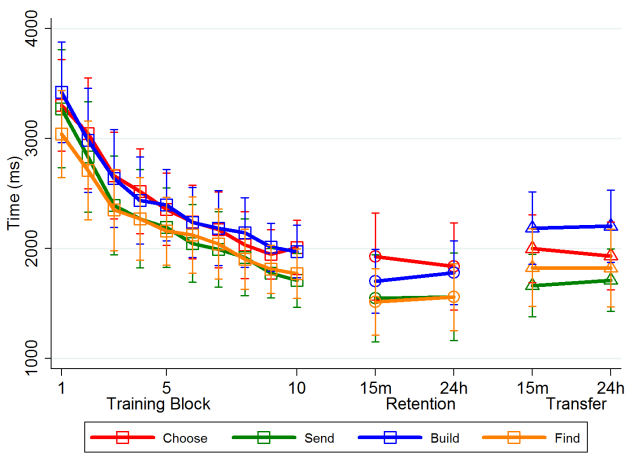

Figure 6: Duration of stroke during training, retention, and transfer, for each gesture. The 'Send' and 'Find' gestures are performed slightly faster than 'Build' and 'Choose'.

attempted to trace the guide would also deviate from the template. However, unlike with crib notes, participants did not attempt to use the post-trial feedback (KP) to correct their previous errors and learn the correct performance.

It is clear that the adaptive guide provides a balance between initial usability and long-term learning. Implementing the adaptive guide in practice, however, is not necessarily straightforward. In the experiment, the number of trials was fixed, and a simple linear model allowed us to gradually remove the guide. In practice, the user is continuously interacting with an interface for an unknown length of time. A simple way to implement the adaptive guide would be to monitor the number of times each gesture was accessed, and provide less guidance each time the guide is accessed, up to some pre-determined threshold.

Other approaches to achieving 'adaptive' style guides are possible as well. One simple way would be to add an access cost to the guide (e.g., a delay [4]), and let users self-regulate the appearance of the guide. This approach was tried in pilot studies, but did not have the desired effect. Many users simply accepted the delay as an intrinsic cost of using the interface, even when the delay was long (e.g., over 1 second). This behavior was present even through it was clearly explained that they did not have to wait for, and use the guide.

In contrast to previous studies, the increased learning observed with crib-notes or the adaptive guides cannot be attributed to increased effort. In fact, the crib-notes guide requires the least effort, as drawing the gesture does not require careful tracing. This was evident with the duration data, which showed the crib-notes had marginally faster input time during training, especially in the earlier trials. If effort were the major determinant of learning, one would expect the dynamic-tracing guide to have the best learning outcomes, as it requires the most cognitive effort (and time) to follow the constantly updating trajectory.

Overall, the type of gesture guide has little impact on duration during training, retention, or transfer (see Figure 4). This is consistent with Bau and Mackay's work that found no difference in the input time between a help menu and a dynamic guide [4]. Participants became faster over time as they became more familiar with the gestures. This is an important find, as it seems to violate the speed-accuracy tradeoff as long as the guide is available. That is, accuracy during training using traceable guides is substantially greater than crib-notes, but the speed of execution is similar, particularly after the first few training blocks.

With these results, it is important to recognize that not all gesture-based interactions target expert-level skill acquisition. Many interfaces are used on an infrequent or casual basis, where the user is not expected to perform at maximum efficiency. For these situations, heavy guidance (with ease of use) is more desirable than limited guidance (with better learnability). However, if efficient expert use is of concern, one may consider a type of guide that initially leads to lower performance, but increases learning and speeds the progression from novice to expert.

Evaluation Paradigm

During retention and transfer, the performance of participants trained with crib-notes was significantly better than the other participants. This demonstrates a severe limitation of the training-only evaluation methods, as the performance changed drastically when guidance was removed. If only the performance data were considered, as in previous studies, one would have reported that the traceable guides were superior with a very large effect size of $\omega^2 = 0.45$. However, when looking at the retention data to assess the learning that occurred, crib-notes proved more effective.

In addition to the focus on learning, another important difference of this work is the use of gesture accuracy as the outcome measure. Previous evaluations tend to use recall as the primary measure [4, 42]. However, as gesture sets become more complex, with a variety of hand shapes and strokes, the ability to articulate precise movements will be a very important measure of user proficiency. Additionally, analyzing the execution of a gesture allows performance improvements to be seen for each block, providing a more detailed look at how users improve with each system.

The performance of nearly all participants suffered after a 24 hour break, with participants that used the traceable guides suffering the greatest performance loss. In general, this indicates users were 'forgetting' how to precisely replicate the gesture. When learning occurs, these losses are much less dramatic. These findings demonstrate the utility of using delayed retention and transfer, as the separation between participants who learned and those who did not becomes greater after a night of sleep [29]. This makes it easier to pinpoint the factors influencing learning.

This study has important implications for the evaluation of future gesture-based interfaces or interactions involving relatively complex movements. When reporting on the effectiveness of various gesture guides, it is imperative that learning be properly assessed, so that designers are aware of the implications (both short and long term) of using an interface. While this places an additional burden on the investigator, it is critical when making claims about the learnability of these interfaces.

This study is limited in the small number of gestures that were studied, and the complexity of the gestures. While a real-world study involving the learning of dozens of gestures would provide more valid results, such a study is impractical to conduct. That being said, it is likely that the results of this study will hold even with more complex 2D stroke gestures. From the training data, performance plateaus after the 5th or 6th block, so participants are already 'overlearning' the gestures to a degree, yet the effect is still clear. Secondly, the guidance hypothesis that underlies the results has been demonstrated in a variety of tasks (including timing, force production, and figure drawing). While this does not guarantee that it will extend to more complex gestures, there is nothing to indicate the contrary. Lastly,

the gestures were chosen to represent a wide range of potential motor actions. It is not clear, however, that the results would generate to a 3D stroke gesture scenario, where the task is fundamentally more complex and would benefit more from guidance [36].

It is also important to note that experts do not tend to precisely reproduce the template. In general, they make simplifications that deviate from the template and allow them to produce the gesture more efficiently, but still be recognizable to the system. While this behavior would not be accurately represented by the strict RMSE measure, our experiments were focused on the reproduction of specific gestures. Similar to the use of a guide that does not adapt to scale or orientation, this simplification allowed us to study the effect of guidance in a controlled fashion. While we anticipate that expert gestures will show more error in a real-world scenario, we do not believe that will negate the guidance hypothesis as it applies to gesture learning.

CONCLUSIONS

With gesture systems that target efficient, expert-usage, it is critical to consider the long-term learning of gestures, and the effects that guidance can have on both performance and learning. Using a novel 'adaptive' guide, we demonstrated that detrimental effects of heavy guidance can be overcome, and that gesture guides can produce a smooth transition from novice to expert. We have also shown that traditional evaluation techniques can be ineffective in measuring learning itself, and that gesture guides may have unforeseen consequences. We have used an established technique from the area of motor learning to evaluate a representative set of gesture guides. We hope researchers and practitioners will use these results in the design and evaluation of future gestural interfaces.

This study provides interesting avenues for future work. The first is to extend this work to larger gesture sets, in a more ecologically-valid scenario. This study made several simplifying assumptions, especially with respect to the small number of gestures. While it is not anticipated that there will be a large difference in the effect of guidance when using more gestures, it is of interest to ensure that the results found in this study do generalize. We have been unifying and standardizing evaluation methods for both the usability and learnability (both the cognitive and motor components) of a gesture system. Until there are clear guidelines for assessing gestural interfaces, manufacturers of commercial systems as well as researchers will continue to develop various ad-hoc methods to assess their systems, making it difficult to compare results across systems and studies.

ACKNOWLEDGEMENTS

The authors wish to thank the reviewers for their helpful comments. This work was supported by Alberta Innovates: Technology Futures, the National Science and Engineering Council, and the Canadian Institutes of Health Research.

REFERENCES

1. Albaret, J. M. and Thon, B. Differential effects of task complexity on contextual interference in a drawing task. *Acta Psychologia*, 100(1-2), 1998, pp. 9-24.
2. Appert, C. and Bau, O. Scale Detection for a priori Gesture Recognition. In *Proc. ACM CHI 2010*. pp. 879-882.
3. Appert C. and Zhai, S. Using strokes as command shortcuts: cognitive benefits and toolkit support. In *Proc. ACM CHI 2009*, pp. 2289-2298.
4. Bau O. and Mackay W. OctoPocus: a dynamic guide for learning gesture-based command sets. In *Proc. ACM UIST 2009*, pp. 37-46.
5. Bennett, M. McCarthy, K., O'Modhrain, S. and Smyth, B. Simpleflow: enhancing gestural interaction with gesture prediction, abbreviation and autocompletion. In *Proc. INTERACT 2011*, pp. 591-608.
6. Bragdon, A., Uguray, A., Wigdor, D., Anagnostopoulos, S., Zeleznik, R., and Feman, R. Gesture play: motivating online gesture learning with fun, positive reinforcement and physical metaphors. In *Proc. ACM ITS 2010*, pp. 39-48.
7. Bragdon, A., Zeleznik, R., Williamson, B., Miller, T., and LaViola J. J. GestureBar: improving the approachability of gesture-based interfaces. In *Proc. ACM UIST 2008,* pp. 2269-2278.
8. Brydges, R., Carnahan, H., Backstein D. and Dubrowski, A. Application of Motor Learning Principles to Complex Surgical Tasks: Searching for the Optimal Practice Schedule. *Journal of Motor Behavior*, 39(1), 2007, pp. 40-48.
9. Cao, X. and Zhai, S. Modeling human performance of pen stroke gestures. In *Proc. ACM CHI 2007*, pp. 1495-1504.
10. Chapman, C. S., Gallivan, J. P., Wood, D. W., Milne, J.L., Culham, J. C. and Goodale, M.A. Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, 2010, 116(2), pp. 168-176.
11. Cockburn, A., Kristensson, P., Alexander, J. and Zhai, S. Hard lessons: effort-inducing interfaces benefit spatial learning. In *Proc. ACM CHI 2007,* pp. 1571-1580.
12. Fitts, P. M. and Posner, M. I. *Human Performance*. Brooks and Cole, Oxford, England, 1967.
13. Freeman, D., Benko, H., Morris, M. R. and Wigdor, D. ShadowGuides: Visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proc. of ACM ITS 2009*, pp. 165-172.
14. Grossman, T., Dragicevic, P. and Balakrishnan, R. Strategies for accelerating online learning of hotkeys. In *Proc. ACM CHI 2007,* 1591-1600.
15. Helsen, W. F., Hodges, N. J., Van Winckel, J. and Starkes, J. L. The roles of talent, physical precocity and

practive in the development of soccer expertise. *Journal of Sports Sciences*, 18(9), 2000, pp. 727-736.

16. Kristensson, P. O. and Denby, L. C. Continuous recognition and visualization of pen strokes and touch-screen gestures. In *Proc. ACM Eurographics SBIM 2011*, pp. 95-102.
17. Kristensson, P. O. and Zhai, S. SHARK: a large vocabulary shorthand writing system for pen-based computers. In *Proc. ACM UIST 2004*, pp. 43-52.
18. Kurtenbach, G. P., Sellen, A. J., and Buxton, W. A. S. An empirical evaluation of some articulatory and cognitive aspects of marking menus. *Journal of Human Computer Interaction,* 8(1), 1993, pp. 1-23.
19. Kurtenbach, G., Moran, T. P., and Buxton, W. Contextual animation of gestural commands. *Computer Graphics Forum,* 13(5), 1994, pp. 305-314.
20. Lai, Q. and Shea, C. H. The role of reduced frequency of knowledge of results during constant practice. *Research Quarterly for Exercise and Sport*, 70, 1999, pp. 33-40.
21. Lepinski, J., Grossman, T., and Fitzmaurice, G. The design and evaluation of multitouch marking menus. In *Proc. CHI 2010*, pp. 2233-2242.
22. Oldfield, R. C. The assessment and ~~analysis~~ of handedness: The Edinburgh inventory. *Neuropsychologia,* 9(1), 1971, pp. 97-113.
23. Ouyang, T. and Li, Y. Bootstrapping Personal Gesture Shortcuts with the Wisdom of the Crowd and Handwriting Recognition. In *Proc. CHI 2012*, pp. 2895-2904.
24. Park, J., Shea, C. H. and Wright, D. L. Reduced frequency concurrent and terminal feedback: A test of the guidance hypothesis. *Journal of Motor Behavior*, 32, 2000, pp. 278-296.
25. Panzer S., Krueger, M., Muehlbauer, T. and Shea, C. H. Asymmetric effector transfer of complex movement sequences. *Human Movement Science*, 2010 29(1), pp. 62-72.
26. Rappin, N., Guzdial, M. Realff, M., and Ludovice, P. Balancing usability and learning in an interface. *In Proc. ACM CHI 1997*, pp. 470-486.
27. Rubine, D. Specifying gestures by example. In *SIGGRAPH 1991*, pp. 239-337.
28. Sainburg, R. L. and Wang, J. Interlimb transfer of visuomotor rotations: independence of direction and final position information. *Experimental Brain Research*, 2002, 145(4), pp. 437-447.
29. Savion-Lemieux, T. and Penhune, V. The effects of practice and delay on motor skill learning and retention. *Experimental Brain Research*, 161(4), 2005, pp. 423-431.
30. Schmidt, R. A. Frequent augmented feedback can degrade learning: Evidence and interpretations. In J. Requin and G. E. Stelmach (Eds.), *Tutorials in motor neuroscience*, 1991, pp. 59-75.
31. Schmidt, R. Motor Learning Concepts and Research Methods. In *Motor Control and Learning: A Behavioral Emphasis,* 2011, pp. 325-346.
32. Schmidt, R. A. and Bjork, R. A. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science,* 3, 1992, pp. 207-217.
33. Schmidt, R. A. and Wulf, G. Continuous concurrent feedback degrades skill learning: Implications for training and simulation. *Human Factors*, 39, 1997, pp. 509-525.
34. Schmidt, R. A., Young, D. E., Swinnen, S. and Shapiro, D. E. Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 1989, pp. 352-359.
35. Shea, J. B. and Morgan, R. Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, 5(2), 1979, pp. 179-187.
36. Sigrist, R. Rauter, G., Riener, R. and Wolf, P. Augmented visual, auditory, haptic and multimodal feedback in motor learning: A review. *Psychonomic Bulletin and Review*, 2012, pp. 1-33.
37. van Nimwegen, C. and van Oostendorp, H. Guidance in the interface and transfer of task performance. In *Proc. ACM ECCE,* 2007, pp. 225-232.
38. Winstein, C. J. and Schmidt, R. A. Reduced frequency of knowledge of results enhances motor skill learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16, 1990, pp. 677-691.
39. Wobbrock, J. O., Morris, M. R. and Wilson, A. D. User-defined gestures for surface computing. In *Proc. ACM UIST 2003,* pp. 193-202.
40. Wobbrock, J., Wilson, A. and Li, Y. Gestures without libraries, toolkits, or training: a $1 recognizer for user interface prototypes. In *Proc. ACM UIST 2007*, pp. 159-168.
41. Zajicek, M. Aspects of HCI research for older people. *Universal Access in the Information Society*. 5(3), 2006, pp. 279-286.
42. Zhai, S. and Kristensson P.O. Modeling Human Performance of Pen Stroke Gestures. In *Proc. ACM CHI 2003,* pp. 97-104.
43. Zhao, S. and Balakrishnan, R. Simple vs. computing mark hierarchical marking menus. In *Proc. ACM UIST 2004*, pp. 33-42.