Investigating Cross-Modal Approaches for Evaluating Error Acceptability of a Recognition-Based Input Technique

JAY HENDERSON, University of Waterloo, Canada
TANYA R. JONKER, Reality Labs Research, USA
EDWARD LANK, University of Waterloo, Canada
DANIEL WIGDOR, Reality Labs Research, Canada and University of Toronto, Canada
BEN LAFRENIERE, Reality Labs Research, Canada

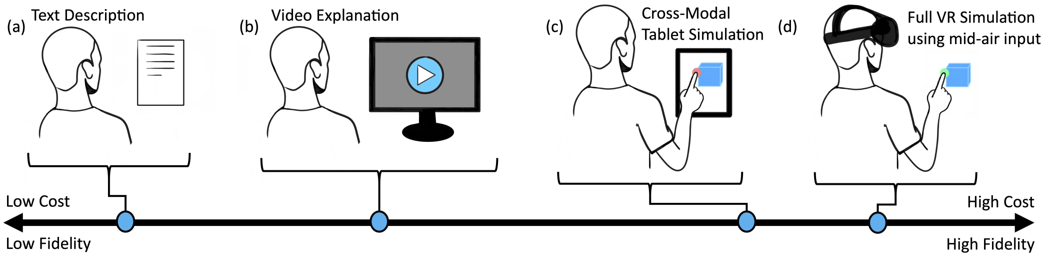

Fig. 1. Example approaches for evaluating error acceptability for a mid-air gestural VR input technique. (a) Text Description, (b) Video Explanation, (c) Cross-Modal Simulation on a multi-touch tablet, (d) Full Simulation in VR.

Emerging input techniques that rely on sensing and recognition can misinterpret a user's intention, resulting in errors and, potentially, a negative user experience. To enhance the development of such input techniques, it is valuable to understand implications of these errors, but they can very costly to simulate. Through two controlled experiments, this work explores various low-cost methods for evaluating error acceptability of freehand mid-air gestural input in virtual reality. Using a gesture-driven game and a drawing application, the first experiment elicited error characteristics through text descriptions, video demonstrations, and a touchscreen-based interactive simulation. The results revealed that video effectively conveyed the dynamics of errors, whereas the interactive modalities effectively reproduced the user experience of effort and frustration. The second experiment contrasts the interactive touchscreen simulation with the target modality – a full VR simulation – and highlights the relative costs and benefits for assessment in an alternative, but still interactive, modality. These findings introduce a spectrum of low-cost methods for evaluating recognition-based errors in VR and a series of characteristics that can be understood in each.

CCS Concepts: • **Human-centered computing** → **User studies**.

Authors' addresses: Jay Henderson, jay.henderson@uwaterloo.ca, University of Waterloo, Waterloo, Canada; Tanya R. Jonker, tanya.jonker@fb.com, Reality Labs Research, Seattle, USA; Edward Lank, lank@uwaterloo.ca, University of Waterloo, Waterloo, Canada; Daniel Wigdor, dwigdor@fb.com, Reality Labs Research, Toronto, Canada, University of Toronto, Toronto, Canada; Ben Lafreniere, benlafreniere@fb.com, Reality Labs Research, Toronto, Canada.

2474-9567/2022/3-ART15 $15.00
https://doi.org/10.1145/3517262

Additional Key Words and Phrases: Acceptability of accuracy; error perception; input techniques

ACM Reference Format:
Jay Henderson, Tanya R. Jonker, Edward Lank, Daniel Wigdor, and Ben Lafreniere. 2022. Investigating Cross-Modal Approaches for Evaluating Error Acceptability of a Recognition-Based Input Technique. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 15 (March 2022), 24 pages. https://doi.org/10.1145/3517262

1 INTRODUCTION

As computing is increasingly embedded in our surroundings, how we interact with computing continues to evolve [9, 28, 42]. Rather than providing deterministic input through a dedicated device (e.g. a mouse, a touchscreen, a keyboard), we may provide input through modalities such as gesture or speech, which offer the promise of a more natural way of interacting, but also push the limits of a system's ability to sense, recognize, and dispatch input. For instance, did the user gesture to provide input, or was it simply natural gesticulation during speech?

Despite significant advances in sensing and recognition, systems that incorporate input techniques based on these approaches – i.e., *recognition-based* input modalities – will inevitably make errors, in which they fail to capture or properly interpret user input. To inform the design of systems that incorporate recognition-based input modalities, it is critical for designers and researchers to understand what the *acceptable error characteristics* are for such input – i.e., what levels of accuracy, precision, and recall are needed to provide a good user experience. Eliciting acceptable error characteristics is a particular challenge for input techniques designed to support ubiquitous computing applications, where the impact of errors may be tied to the specific application and context in which a technique is used. As an example, errors in mid-air gesture input to support augmented reality in a relatively static environment can be simulated and evaluated using room-scale 3D motion capture; however, the same input modality, used in a mobile context using smart glasses, may introduce new considerations that change the acceptable error characteristics of the system. Testing new applications and scenarios, therefore, requires building additional simulated environments, which can be costly. Ideally, we would like to be able to infer aspects of error characteristics early in development to limit the cost and effort spent on development, user testing, and improving sensing and recognition techniques.

With the exception of actually building and testing a functional prototype, we are aware of no approach that directly explores alternate mechanisms for eliciting acceptable error characteristics in recognition-based gestural input. However, in a related domain, Kay et al. developed a text-based survey tool to determine the error characteristics that users find acceptable for ubiquitous computing applications that depend on basic sensing and machine learning classifiers [25]. Scenarios explored included, for example, a home alarm system with false alarms, and eliciting from users acceptable error characteristics for cases where the alarm system called the police when triggered vs. simply notifying the user via text message. However, it is unclear if Kay et al.'s approach can be extended to input techniques, which differ in several ways. First, in recognition-based input techniques such as gesture or speech, input is provided through explicit actions of the user (versus implicit input through sensing in the applications studied by Kay et al., with the user simply responding to inferences made by the system). As well, the dynamics of an input technique may be more difficult to express through text descriptions alone [14]. Finally, the acceptability of errors with recognition-based input techniques may depend on a user's hands-on experience with failures, and a text description has limited ability to communicate the practical experience of use and error.

This paper seeks to investigate a broader design space of low-cost methods for eliciting acceptable error characteristics, including supplementing text with video demonstrations of accuracy scenarios, and a *cross-modal acceptability elicitation* approach that uses a well-established input modality (e.g., multi-touch input) to simulate an input technique that is under development. The cross-modal approach was motivated by the observation that many of the essential characteristics of an input technique that may influence the acceptability of errors, such as timing and the type of feedback available (e.g., audio, visual, or haptic) can be simulated by existing input modalities.

To understand how low-cost approaches and widely-used input modalities can provide insights into the development of novel input techniques, mid-air freehand gestural input in VR was chosen as an example target modality, and acceptable error characteristics were elicited using three alternative modalities (text descriptions, as in Kay et al. [25]; video demonstrations; and a cross-modal approach via a multi-touch tablet). Two application scenarios were tested: a gesture-driven infinite running game and a painting application.

Our study results establish that text descriptions are insufficient on their own to elicit acceptable error characteristics for our tested target modality. Video can help clarify the dynamics of input interactions and types of errors that can occur, and helps with understanding application contexts, but interactive approaches (such as the cross-modal approach) are required to effectively convey the effort and frustration that comes from errors, which are key to understanding error acceptability. Overall, this work contributes the idea of developing low-cost approaches for eliciting acceptable error characteristics for input techniques, and presents empirical evidence to highlight the relative benefits and drawbacks of text, video, and cross-modal approaches for this purpose.

2 BACKGROUND AND RELATED WORK

This exploration into user error acceptability is related to, and inspired by, research on error acceptability, prototyping methods, and inter-modal learning and transfer.

2.1 Understanding Error Acceptability

Input techniques that rely on recognition are susceptible to two types of system errors. With *false positives*, a system believes that a user has generated input (e.g., a gesture) when they have not. With *false negatives*, a user has generated input but the system fails to recognize this input. The error characteristics of these systems are typically expressed in terms of *precision* (i.e., the fraction of instances where a recognizer has detected input and the recognizer is correct) and *recall* (i.e., the fraction of instances where input from a user is correctly detected by the recognizer). Expressed as equations in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN):

$$P = \frac{TP}{TP + FP} \quad \text{(precision)} \qquad R = \frac{TP}{TP + FN} \quad \text{(recall)}$$

Prior research on input techniques has largely focused on reducing errors by developing more accurate recognition algorithms [32, 36, 39], feedback mechanisms [3, 21], or techniques that help users to learn how to perform gestures [3, 14]. There has been comparatively little work to understand the user experience of different error types, or how to find acceptable error characteristics for particular applications. Though we are unaware of any systematic approaches to evaluating acceptable error characteristics for input techniques, recent work has demonstrated that users prefer false negative errors over false positive errors [29], and that experiencing multiple errors in quick succession has a particularly negative effect on user experience [23].

Kay et al. called out the need for a method to understand which error characteristics are tolerable to users of ubiquitous computing systems that are driven by simple sensors and machine learning classifiers [25]. They developed a survey tool and *acceptability of accuracy* metric to answer this question. Potential users were asked to imagine a system based on a short text description, and then rate a set of accuracy scenarios describing the error characteristics of a classifier underlying the system. A model of acceptability of accuracy was fit to the responses, and captured the relative weight that users placed on precision versus recall – used to evaluate a classifier that was under development.

Kay et al.'s approach considered applications at a high level, i.e., at a granularity that could be described in a short paragraph of text, including descriptions of the consequences of errors. This is well suited to systems with simple interfaces, or early stage design where the finer details of interface and presentation have yet to be worked out. However, prior work has suggested that the details of how intelligent features are integrated into an interface can influence error acceptability. For example, Roy et al. demonstrated that users can tolerate lower accuracy in exchange for greater control in intelligent assistance systems [35]. The intelligibility and interpretability of intelligent features have also been shown to influence error tolerance [10, 41]. Finally, recent work by Kocielnik et al. has demonstrated techniques to mitigate the impacts of error by setting a user's expectations [27].

The present work is inspired by Kay et al.'s approach, and the value of understanding which error characteristics are acceptable to users early in the design process. However, it seeks to expand the general approach into a new domain (i.e., input techniques), and considers a broader design space of techniques for evaluating the acceptability of accuracy in potential applications, including video and interactive demonstrations.

2.2 Prototyping Methods

Many prototyping tools and techniques [4] and discount usability engineering methods [12, 13] have been developed to ease the challenges inherent in early design processes, including Wizard of Oz techniques [6], low-fidelity prototyping methods [11], and methods using techniques such as storytelling, video, and stagecraft [5, 34, 45]. The present exploration draws inspiration from these techniques, and their demonstration of the value in evaluating early representations of a system that is being designed, but applies this general approach to the distinct problem of assessing acceptable error characteristics.

A related thread of prototyping research has examined how to facilitate designers or researchers in rapidly building interactive prototypes, for example through hardware toolkits [16], prototyping platforms [19, 31], or leveraging available input/output technology [30, 37]. In principle, these methods could be used to generate representations of an input technique for the purpose of evaluating acceptable error characteristics, however, they have yet to be applied specifically for this task. Likewise, commercially available prototyping tools such as Adobe XD and Sketch may be suitable for this use – though, in practice, these tools are focused on expressing and enabling early evaluation of design elements, rather than the finer-grained interaction dynamics which may play a role in error acceptability.

2.3 Inter-modal Learning and Transfer

Prior work has also investigated how systems can use alternative input modalities to teach users skills in a target modality. For example, instructional videos, iconic representations, and interactive guidance are popular techniques that have been used to instruct a user how to perform spatial, 3D, or mid-air interactions [1, 8, 15, 20, 22, 38]. In particular, skill with directional surface gestures, an alternative modality, has been shown to effectively transfer to in-air gestures [20]. The present exploration investigates the degree to which similar alternative modalities can be used to understand a target modality, but the objective is understanding the acceptable error characteristics for the target modality, rather than skill transfer.

3 MODALITIES FOR ERROR ACCEPTABILITY ELICITATION

The objective of this work is to investigate the use of alternative modalities – text, video, or cross-modal – for eliciting acceptable error characteristics for an input technique. One way to think of the space of potential modalities is in terms of a trade-off between the *fidelity*, with which a modality represents an input technique being tested, and the *cost* involved in producing such a representation. In general, we would expect that higher fidelity representations of a technique being tested will be better for eliciting feedback that is valid for that technique as well. However, we would also expect that higher fidelity representations will come with higher costs to develop. In this framing, text descriptions and a full simulation of an input technique can be seen as two end-points of a continuum, with the former optimizing cost, and the latter optimizing fidelity (Figure 2).

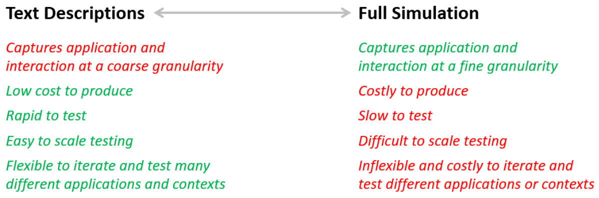
Text Descriptions
Full Simulation
Captures application and interaction at a coarse granularity
Low cost to produce
Rapid to test
Easy to scale testing
Flexible to iterate and test many different applications and contexts
Captures application and interaction at a fine granularity
Costly to produce
Slow to test
Difficult to scale testing
Inflexible and costly to iterate and test different applications or contexts

Fig. 2. Text descriptions and a full simulation represent a continuum in terms of trading off cost and fidelity.

A key insight which motivated our current investigation is that, for the purposes of eliciting acceptability of errors, the fidelity of a representation may only matter to the extent that it does or does not convey aspects of the input technique that influence the acceptability of errors. Thus, there may exist modalities that can be used to represent an input technique at relatively low cost, while still being useful for eliciting acceptable error characteristics. In the next section, we discuss several attributes of input techniques that we believe have the potential to influence error acceptability, which guided our choice of modalities to test.

3.1 Input Technique Characteristics

Several aspects of input techniques seem likely to influence the acceptability of errors:

Application and Usage Context – Input techniques are used as a means of providing input to applications, and the specific application and context in which input is provided are likely to influence the consequences of errors and their effect on user experience. For example, a false positive error that triggers a 'send message' action prematurely in an email application could have highly negative consequences, whereas a similar error with effects that are easily reversible may not.

Timing, Responsiveness, and Feedback – The timing, responsiveness, and feedback provided by an input technique is also likely to influence the experience of errors. For example, a false negative error when performing a gesture in a technique with highly-responsive sensing may be less frustrating, because the gesture is quick to perform and quick to retry. In contrast, a technique with high latency may make such errors more frustrating.

Experience over Time – The occurrence of one error with an input technique may seem inconsequential, but the experience of many such errors may add up to a negative experience. Thus, for input techniques, error acceptability may not be apparent from a description or demonstration of a single error instance, but may require an approach that allows a user to experience the rate of occurrence of errors over time.

The above set of characteristics suggest that modalities that enable the user to try out an input technique, or at least observe the technique being used, will have advantages for eliciting acceptable error characteristics. This hypothesis is tested in the studies reported in this paper.

3.2 Cost Characteristics

Modalities may also differ in the costs that they impose on a researcher attempting to use the modality to represent an input technique being tested. First, there is the *cost to produce* the representation of the technique, which may include writing descriptions, development time, hardware setup and prototyping, among other factors. Second, there is the *cost to test*, or the time it takes to elicit information from potential users. For example, the cost to test would be higher for a modality where users must spend time trying out an input technique to experience errors as they occur, versus one where they simply read a description of that technique. Related to cost to test is *scalability* – some techniques may be easy to distribute remotely to many potential users, enabling rapid testing, whereas others may require custom hardware setups that permit only one person to experience a technique at a time. Finally, particular modalities may vary in their *flexibility*, or how easily they can be modified or adapted to test variations of a technique, or to test the technique in different scenarios or environments.

The next section considers potential low-cost modalities in terms of the input technique characteristics they could potentially express, and the cost characteristics discussed in this section.

3.3 Potential Low-Cost Modalities

Three modalities were chosen for testing that seemed, in-principle, capable of capturing the input technique characteristics discussed above, while also representing a range of potential costs. Specifically, video demonstrations and cross-modal interactive demos are considered. Text descriptions and full simulations of the target modality are briefly discussed as well, as they represent endpoints for the fidelity-cost continuum.

3.3.1 Text Descriptions. At the low end of both fidelity and cost are text descriptions, which ask potential users to imagine interactions with an input technique, and errors that may occur. This has the benefit of a low cost to produce and test, and high scalability and flexibility. A potential pitfall of this approach is that text descriptions cannot capture most of the input characteristics discussed in the previous sections, and users may interpret text descriptions in different ways, which could influence their assessment of error characteristics.

3.3.2 Video Demonstrations. Video demonstrations are often used to demonstrate interaction techniques [1, 8, 15, 22, 38], and one might imagine them being used to supplement text descriptions and visually communicate timing, responsiveness, feedback, and the various consequences of errors in applications. This modality has a higher cost to produce than text descriptions, but could still be achieved relatively cheaply, as an input technique does not actually have to be implemented, but could be merely simulated for a video. While it may take somewhat longer to view videos than to read text, the cost to test, scalability, and flexibility of this approach are close to that of text descriptions.

Although video can easily convey visual details to the user, simply viewing an interaction technique in use may not be sufficient to convey the experience of using it. As well, some forms of feedback cannot be communicated through video (e.g., haptics), and the perception of the ease of performing input may be skewed based on the representation in the video.

3.3.3 Cross-Modal Interactive Demos. The use of an established input modality to provide an interactive demonstration of the target input technique could be useful because most of the elements of the target modality could be preserved, including timing, feedback, the experience of errors and their consequences both in the moment and over time. Moreover, by choosing an established modality with high accuracy (e.g., multi-touch input on touchscreens), errors can be selectively injected to degrade the simulated accuracy and create any target set of error characteristics. Cross-modal approaches with widely-available input techniques could also permit remote or crowd-sourced user studies which would not be possible with specialized prototype hardware – i.e., the approach has some advantages in terms of scalability.

There are, however, some potential disadvantages to this approach. First, the cost to produce is higher for an interactive demo, and it may be more costly to change (i.e., reduced flexibility), although well-established modalities, such as touch-screen devices have the benefit of mature development platforms and ecosystems that can be used for prototyping or developing interactive experiences (e.g., Adobe XD, Unity 3D). The cost to test will also be greater, since the user must interact with the cross-modal demo to experience the error characteristics. Testing using a well established modality may also impose expectations that bias the evaluation of the simulation of the target modality, e.g., simulating an input technique with pointing uncertainty using a mouse, which is known to have high accuracy. Feedback mechanisms may also exist in a simulation modality that do not exist in the target modality, e.g., if a mouse drag is used to simulate a mid-air gesture, there is explicit tactile feedback when the mouse button is pressed that will not be present in a mid-air gesture. Despite the above challenges, we see cross-modal interactive demos as a potentially promising approach, for their ability to represent the interactivity of the target technique.

3.3.4 Target Modality Simulations. The highest fidelity approach, is to create an accurate simulation of the target modality, such that the desired range of potential error characteristics can still be simulated. For example, to simulate mid-air gesture detection for camera-based hand tracking, a data glove could be used that provides more accurate sensing than can be achieved using the current camera-based tracking techniques. In many cases, highly accurate simulations may be an ideal approach, but they have the disadvantages of being costly to produce and having low scalability and flexibility, because they often require working with custom hardware or prototype systems that are difficult to develop for, or which require infrastructure to test (e.g., room-sized motion tracking systems).

3.4 Practical Importance of Eliciting Acceptable Error Characteristics

Unless we can somehow explore acceptable error characteristics, system design and development must proceed without an understanding of what sufficient performance might be. If we can, as a first step, explore the impact of error characteristics (e.g., how variations in precision and recall influence user experience), then two options present themselves to designers: they can invest resources in producing systems of sufficient precision and recall to satisfy user requirements; or they can explore ways to manage error intelligently so that attainable error characteristics will be sufficient to support a modified form of interaction. In either case, low-cost mechanisms for understanding error characteristics *a priori* provide a significant advantage.

3.5 Section Summary

In this section we have presented a set of modalities that could potentially be used to elicit error characteristics. Specifically, while past research has presented some indirect evidence of the utility of text-based descriptions [25], we note potential challenges with textual descriptions for recognition-based input techniques and propose alternatives including video demonstrations and cross-modal interactive demos. Both of these alternative modalities fall short of the requirements of a full simulation in the target modality, but may provide valuable insight into error characteristics at a much lower cost than a full simulation, which could justify their use at varying stages of the development process for an emerging input technique.

To evaluate the effectiveness of these alternative modalities for eliciting acceptable error characteristics, the remainder of this paper presents two studies. First, we contrast text-based descriptions [25] with video demonstrations and cross-modal interactive demos to highlight differences in elicited characteristics. Motivated by the results of our first study showing promise for cross-modal interactive demos, we perform a second study that compares cross-modal interactive demos to a full simulation in the target modality.

4 EXPLORING ALTERNATIVE METHODOLOGIES

To understand how users perceive error characteristics across different modalities (text, video, cross-modal), a user study was conducted evaluating three different modalities and two application scenarios. Mid-air gesture interaction in VR was chosen as a sample target interaction technique because it relies on ambiguous input and sensing.

4.1 Participants

Thirteen participants who had access to an iPad running iOS 9.4 or later, were recruited to participate in the study through word of mouth. Seven participants identified as male, six as female (μ = 32 years, σ = 11.7 years, range = 23 to 57 years). Eleven participants self-reported that they were right handed, one was left handed, and one was ambidextrous. The study took approximately one hour to complete. Participants were provided with a $35 Amazon gift card in local currency as a honorarium for their time.

4.2 Applications and Apparatus

Two applications were created to evaluate the degree to which application demands and context would influence the user acceptability of recognition errors. Motivated by work by Gutwin et al. [17], two applications that leveraged an in-air pinch gesture interaction were implemented. The first application, *Endless Runner*, included a key temporal demand element, which was deemed *high urgency*. The second application, *Finger Painting* had no temporal performance requirement so it was deemed (*low urgency*).

Index finger and thumb pinch gestures were chosen as the target movements because they are commonly performed freehand gestures in novel technologies, such as the Microsoft HoloLens [33], due to their natural delimiters and synonymy to grasping of an object [43]. Though this input technique is thought of as "robust" to recognize [2, 43] via cameras, this is contingent on the user's hand being oriented in a way that can be in full view of the cameras and under optimal lighting conditions. To increase the expressivity for more realistic uses of the pinch gesture, it was paired with left, right, up or down uni-stroke gesture, followed by releasing the pinch, to complete an action. Spatial movements are often coupled with the pinch gesture to complete richer interactions [33].

Both applications were developed using Unity (2019.2.17f1) and ran on iPad devices running iOS 9.4 or higher. iPads were chosen as the input apparatus to ease gesture interaction and to preserve elements including the larger spatial area required for VR approaches. The code for each application was compiled via Xcode (11.4.1) and deployed via TestFlight to participants.

To ensure each scenario received the correct distribution of errors/non-errors, gestures were organized into an array of size of TPs + FPs + FNs. Each index of the array was randomly assigned to either, TP, FP or FN. False negative (FN) errors appeared whenever the system was intending to ignore a correctly performed gesture. To generate realistic false positive (FP) errors, users' index finger movements were tracked and if the user moved past a threshold in a particular direction, an error was injected. Otherwise, an error was injected upon a user pinching to avoid the user completing a TP at an index containing a FP.

4.2.1 Endless Runner. The Endless Runner application began with a red ball displayed on a white track (Figure 3). The goal is to remain on the track for as long as possible without hitting an obstacle (which are introduced at random intervals) or falling off of the track. The user could perform four actions by making a pinch and directional gesture in the desired trajectory (e.g., pinch and up movement jumped over an obstacle, pinch and down shrunk the red ball under an obstacle, pinch and left or right moved the ball left or right to avoid an obstacle). If the ball collided with an obstacle or fell off of the track, there was a five second penalty, after which the game would resume.

4.2.2 Finger Painting. The finger painting application began with a blank canvas in the center of the screen. The goal of the application was for the user to recreate a series of images displayed on the left side of the canvas (Figure 4). The goal images comprised of four colours (i.e., blue, red, green and black). The user could select a colour to paint with by completing a directional gesture in one of four directions: left (green), right (red), up (blue) and down (black). Once the user had completed the image, they could select "next" to load the next image.

4.3 Error Acceptability Evaluation Methodologies

As previously mentioned, we wished to gain insights into the benefits and challenges inherent in various modalities for communicating error characteristics to users. Three of the modalities from the design space were evaluated: text descriptions, video demonstrations, and the cross-modal approach.

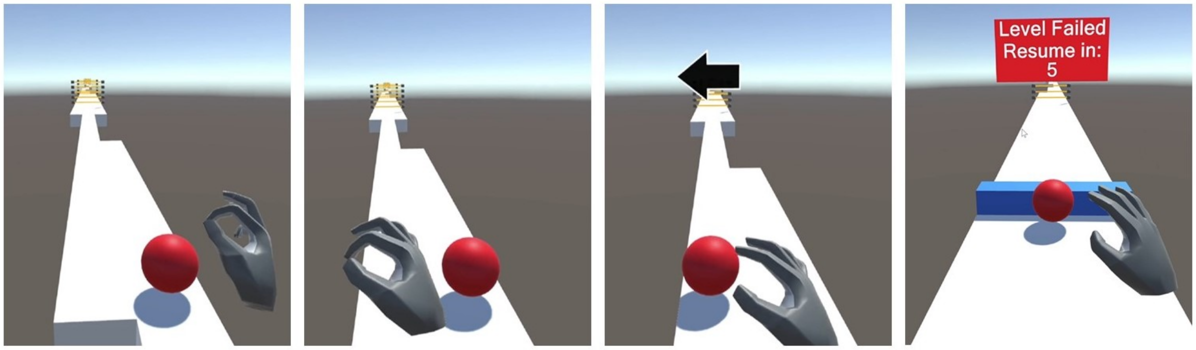

Fig. 3. The Endless Runner application. Left three images: A gesture is used to move the ball to the right; Right-most image: a penalty is incurred for hitting an obstacle.

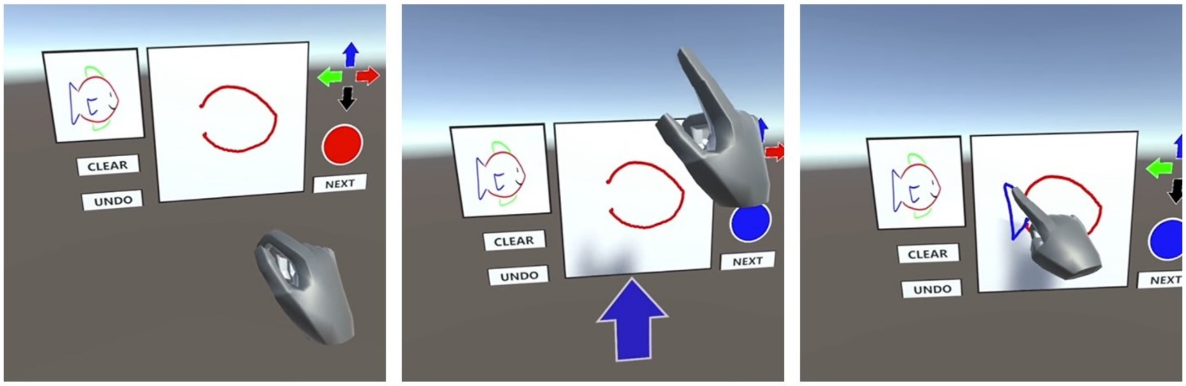

Fig. 4. The Finger Painting application. The user selects blue with a gesture (i.e., pinch-move-up-release) and then paints.

4.3.1 Text Descriptions. To avoid researcher bias, an HCI researcher external to the research team was hired to produce write-ups for each application scenario and accuracy condition. In our request to produce the write ups, we did not reveal the goals or hypotheses of the current project. These scenario descriptions followed the same structure of Kay et al.'s work [25], including the use of the *True Positives* (TP), *False Negatives* (FN), and *False Positives* (FP) values corresponding to P of 0.941176 and R of 0.80. A sample scenario description produced for the *Endless Runner Game* was as follows.

Please imagine the following:

- ***20 times*** *during the game you perform a hand gesture to avoid an obstacle.*
  - ***16 of the 20 times*** *that you perform a gesture, the system (correctly) recognizes the gesture so you avoid the obstacle.*
  - ***4 of the 20 times*** *that you perform a gesture, the system (incorrectly) does not recognize the gesture so you do not avoid the obstacle. You either receive a 5 second penalty, or, you have enough time to perform the gesture again, so you avoid the obstacle and receive no penalty.*
- ***1 other time*** *during the game, the system (incorrectly) recognizes a gesture that you did not perform, causing the ball to jump, shrink, or roll off the track. You receive a 5 second penalty for not avoiding an obstacle, or, no penalty if no obstacle present.*

Table 1. Accuracy scenarios used within the experimental protocol, e.g., P95/R95 indicates 95% precision, 95% recall.

Scenario	TP	FP	FN	P	R
P95/R95	19	1	1	0.95	0.95
P94/R80	16	1	4	0.94	0.80
P83/R95	19	4	1	0.83	0.95
P80/R80	16	4	4	0.80	0.80

4.3.2 Video Demonstration. The video condition consisted of screen recordings of a user completing each scenario in a VR simulation. To ensure participants understand the types of errors and gestures to watch for, prior to showing the error distribution video to participants, a video was shown outlining each possible non-error, error, and consequence. Error consequences were balanced across applications (i.e., the cost of a 5 second penalty would in the Endless Runner application was balanced with a cost to having to erase a stroke in the Finger Paint application). Each video was 1-2 minutes in length.

4.3.3 Cross-Modal Approach. To simulate the mid-air pinch gesture via a lower fidelity interaction method, an application was developed to recognize pinch gesture input via multi-touch input. The pinch gesture interaction worked as previously described, however all gestural input was completed while touching the screen. To begin a pinch movement, the user was required to touch both their index finger and thumb on the screen, pinch both fingers together, and while continuing to maintain contact with the screen, drag their fingers in one of the four cardinal directions. Once completed, the user had to spread their fingers apart or remove them from the screen to conclude the gesture. To control for accuracy expectations using a touch screen, we instructed users that errors were designed to simulate those that may appear in an imperfect mid-air tracking system, so they should imagine they are interacting with such a technique.

4.4 Accuracy Scenarios

The accuracy scenarios that were used in the study were chosen to include distributions that would capture coarse differences in overall error rates (scenarios P95/R95 (95% precision, 95% recall); and P80/R80) and finer grained differences between biasing towards different types of errors (scenarios P94/R80 and P83/R95; Table 1). These rates of precision and recall are higher than those tested in Kay et al.'s work, to reflect a range that is realistic for input techniques.

4.5 Procedure

The experiment used a within-subjects design with three factors: *application* (Endless Runner, Finger Painting) and *modality* (Text, Video, Cross-Modal), and *accuracy scenario* (P95/R95, P94/R80, P83/R95, P80/R80). Each participant experienced each factor and level with modality held constant for each participant (i.e., Text, Video, then Cross-Modal. The ordering of applications was alternated between participants, and accuracy scenarios were randomized by application and modality. After each presented scenario (i.e. error distribution of an application in a particular modality), participants completed a questionnaire. At the conclusion of the study, participants were asked a series of open-ended questions intended to probe differences in the ability of each modality to convey error information to the user.

The experimenter and participants conducted a video call to supervise the study and ensure proper functioning of the study software. The text descriptions, videos, and surveys were all deployed via a web application hosted on the experimenter's computer, which also captured participants' responses.

4.6 Data Collection

To evaluate the acceptability of our input technique at varying levels of precision and recall, five 7-point Likert scale questions were administered to participants for each trial task and condition (1 - strongly disagree, 7 - strongly agree).

Questions 1 and 2 were adapted from Kay et al.'s original paper to apply to input techniques [7, 25, 40]. While question 1 is a relatively straight-forward adaptation, a question asking *"I would find this application to be useful"* would not apply to input techniques. Thus, we attribute the perceived usefulness of an input technique as how useful a particular input technique is in allowing a user to perform well. We did not include Kay et al.'s question 3 [25] as *intent to use* lacked meaning when assigning arbitrary tasks for users to complete.

To evaluate the workload required to perform an input technique, questions 3 and 4 incorporated effort and frustration aspects of the NASA TLX [18]. Question 5 is an overall metric for user experience for the presented technique.

(1) The accuracy of the input technique would be acceptable for this application. (*acceptability of accuracy*)
(2) I would be able to perform well with this input technique. (*perceived usefulness*)
(3) I would have to work hard (mentally and physically) to perform well using this input technique. (*effort*)
(4) I would be frustrated using this input technique. (*frustration*)
(5) I would be satisfied with the overall user experience of this input technique. (overall UX)

120 Likert responses were collected per participant. While transmitting data, partial data was lost for one participant, so their data were removed from analysis. Therefore, twelve participants' responses were included in the analysis.

4.7 Results: Likert Response Data

An align-and-rank transform (ART) [26, 44] factorial ANOVA was run on each of the five Likert scales that were administered (i.e., Acceptability, Usefulness, Effort, Frustration and Overall UX). Significant differences were found for all questions on Modality, Application, and Scenario (Table 2). The sections that follow present post-hoc analyses for each factor. Likert scale results by question are depicted in Figure 5.

4.7.1 Modality. When evaluating the three modalities, post-hoc tests using Tukey's HSD indicated significance (p<.05) for Acceptability, Effort, Frustration, and Overall UX between Cross-Modal vs. Text, and Cross-modal vs. Video, but not for Text vs. Video; and for Usefulness for Cross-Modal vs. Text only. This suggests that the added fidelity of the cross-modal approach led participants to provide different ratings of error acceptability.

4.7.2 Application. As indicated in Table 2, significant effects were found for each question across application, suggesting that the different applications do have different acceptable error characteristics, and they came through in the modalities we tested.

4.7.3 Scenario. When evaluating the four precision and recall scenarios, post-hoc tests using Tukey's HSD indicated significant differences (p<.05) between each pair except for P94/R80 vs. P83/R95 in each Likert question. This provides evidence of a difference in acceptability across accuracy scenarios as well.

4.7.4 Interaction Effects. Looking at interaction effects, we found no significant interactions between Scenario:Modality or Scenario:Modality:App. The remaining interactions are shown in Figure 3.

For Scenario:App, we found significant differences for Usefulness and Overall UX on (P95/R95 - P94/R80 | Game) vs. (P95/R95 - P94/R80 | Paint), and on (P95/R95 - P83/R95 | Game) vs. (P95/R95 - P83/R95 | Paint). For Usefullness, we found differences for (P95/R95 - P80/R80 | Game) vs. (P95/R95 - P80/R80 | Paint). These findings suggest that the degradation in user experience of the input technique in response to error characteristics with

Table 2. Align-and-Rank Tranform RM-ANOVA results for Study One Likert data for individual factors of Scenario, Modality and Application (App).

	Modality			Application			(P,R) Scenario		
	F	df	p	F	df	p	F	df	p
Acceptability	6.597	(2,253)	<.01	50.810	(1,253)	<.001	42.335	(3,253)	<.001
Usefulness	4.292	(2,253)	<.05	68.946	(1,253)	<.001	44.774	(3,253)	<.001
Effort	11.833	(2,253)	<.001	83.777	(1,253)	<.001	19.118	(3,253)	<.001
Frustration	17.399	(2,253)	<.001	83.276	(1,253)	<.001	40.342	(3,253)	<.001
Overall UX	10.290	(2,253)	<.001	54.711	(1,253)	<.001	43.534	(3,253)	<.001

lower precision and/or recall is not equal across the two applications. In particular, follow up analyses show that the degradation is greater for Game than Paint, which makes sense given the added urgency in this application.

For Modality:App, we found significant differences for (Cross-Modal - Text | Paint) vs. (Cross-Modal - Text | Game) for all Likert questions. Additionally, for Usefulness, Effort, Frustration, and Overall UX, we found significant differences for (Cross-Modal - Video | Game) vs. (Cross-Modal - Video | Paint). No significant interactions were found for Text vs. Video. These findings suggest that the added fidelity of the cross-modal approach was more beneficial in revealing the acceptable error characteristics for one application than the other.

Table 3. Align-and-Rank Tranform RM-ANOVA results for Study One Likert data considering interactions of Modality, Application (App), and Scenario. There were no significant interactions between Modality:Scenario or between Modality:Application:Scenario

	Modality:App			App:Scenario		
	F	df	p	F	df	p
Acceptability	3.598	(2,253)	<.05	n.s.	n.s.	n.s.
Usefulness	4.859	(2,253)	<.01	3.229	(3,253)	<.05
Effort	3.388	(2,253)	<.01	n.s.	n.s.	n.s.
Frustration	1.284	(2,253)	<.001	n.s.	n.s.	n.s.
Overall UX	4.284	(2,253)	<.05	3.515	(3,253)	<.05

4.8 Results: Qualitative Responses

Outside of the Likert questions, after the experiment participants were asked if each modality provided sufficient information to answer the questions after each scenario. 7/12 of participants said yes for the Text modality, 11/12 for Video, and 12/12 for Cross-Modal.

For the open-ended questions, participants were asked what was understandable and unclear in each presented modality. Seven participants noted the text descriptions allowed for understanding of the exact numerical error rates, e.g., *"very clear quantitative data"* (P3). However, seven participants also reported that, in general, the descriptions were difficult to understand; due to either lack of visuals (P1, P4, P7, P8) or general readability (P1, P2, P9, P10). Others noted it was challenging to understand effort (P12), how errors would "feel" (P11), or implications for frustration (P1, P6).

For the Video modality, all participants felt that the videos clarified the prior explanations. Two participants stated that the videos gave a complete explanation. Other aspects that were clarified include understanding of the temporal factors (P1, P6, P9, P12) and how to use the techniques (P4, P5, P8). One participant noted the video

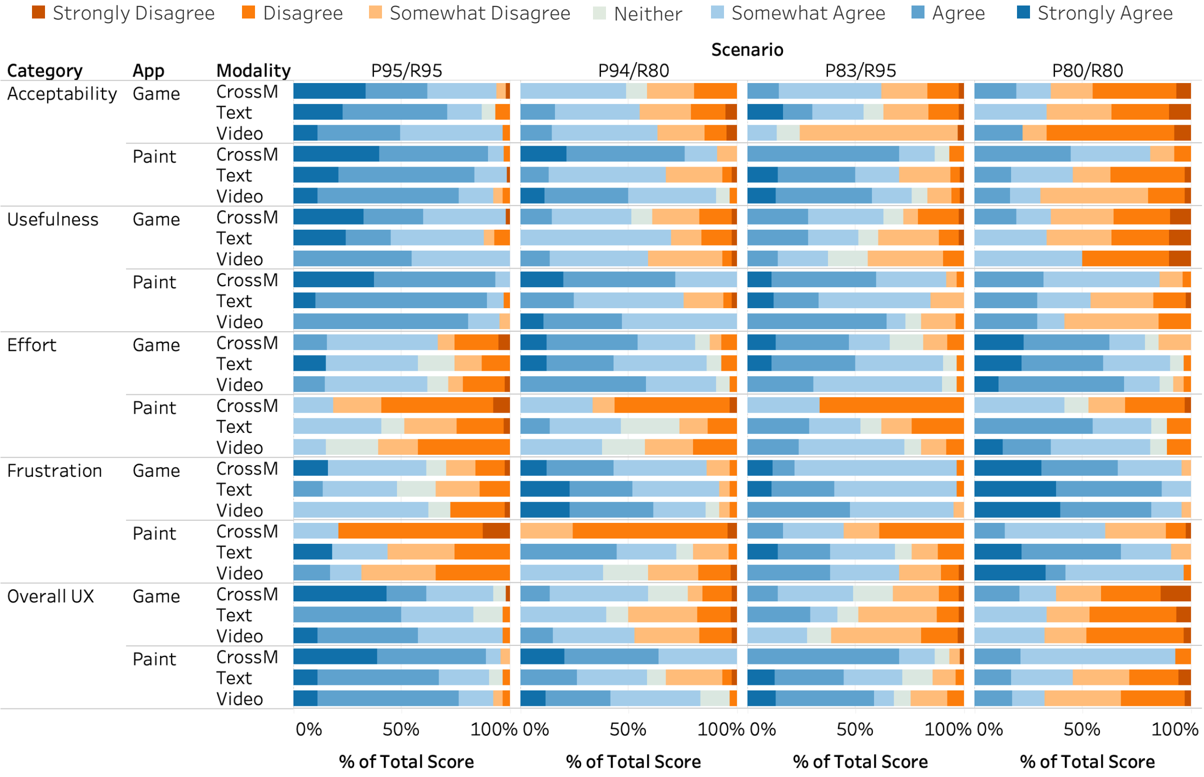

Fig. 5. Summary of Likert data for in Experiment One by Category, Modality, and Application.

helped them *"feel the pain"* of the user (P7). When asked what aspects were lacking clarity, responses included more perceptual and emotional implications, such as frustration (P3, P4), annoyance with the running game (P7), how much effort would be required (P12), and *"the actual experience of making decisions"* (P9). P2 noted it was easy to lose a viewer's attention because the videos were time consuming.

Finally, all but one participant felt the Cross-Modal (tablet) modality improved their understanding of the input technique and characteristics of errors that may appear. Four participants felt that emotional impact became clarified through experience, including frustration (P1, P4, P5, P6), how *"fun"* it would be (P4), and tolerability and acceptability (P5, P7, P11). Actual motor requirements (P3, P4) and effort (P12) were also noted as becoming more understandable. One participant noted that they could get a better understanding of when an error was a false positive vs. false negative (P8). In terms of what participants found to be unclear, one participant noted it was difficult to understand why errors would occur, especially when they weren't touching the iPad (P1). Two participants were unsure what the actual numerical error rate was when using the simulation (P11, P12).

When prompted for additional comments, some participants noted that some differences in the application types could only be communicated via experience, e.g.,*"For both the text and video simulation I assumed that these errors would give me the same response for both applications, however it wasn't until I completed the tablet simulation that I realized the Endless Runner errors were far more annoying"* (P1). Another participant noted that in the finger painting application *"it's a more passive and calming activity, so little mistakes are easier to forgive"*

(P5). In contrast, when considering the actual input technique, one participant responded *"I think the pinch/swipe motion is kind of awkward on a tablet compared to in the VR setting"* (P6).

4.9 Study Synthesis

The findings from this study suggest that text descriptions are insufficient on their own to elicit acceptable error characteristics for a mid-air input technique in virtual reality. This can be observed in the significant differences we observed between Likert responses for the text descriptions and the cross-modal representation of the system, and the finding that only 7/12 participants felt the text descriptions provided sufficient information to answer the Likert questions, versus 11/12 for video and 12/12 for the cross-modal representation.

While participants expressed that the addition of video helped clarify the text descriptions, the lack of significant differences in responses for most Likert questions between the text and video conditions suggests that video is still falling short of the experience provided by the cross-modal approach. This is reinforced by participants' comments that the cross-modal approach further improved their understanding of the technique and error characteristics, and the significant differences we found for the Likert responses between the cross-modal technique as compared to the text and video conditions. Given this, it is interesting that 11/12 participants agreed that the video modality provided sufficient information. It may be that video gives a false sense of confidence, or that participants' responses to this question were inflated because they answered the question at the end of the study, after they had experienced the higher fidelity approaches. If the latter explanation is correct, the response of 7/12 participants that text provided sufficient information may be inflated as well. Finally, the comment by one participant that they had difficulty keeping their attention focused on the videos is interesting, and may indicate a pitfall to using this approach.

Of the modalities tested in this study, the cross-modal approach appeared to be the best in terms of conveying error characteristics. In particular, participants' feedback suggests that the cross-modal approach was effective at conveying how differences in application influenced the impact of errors (e.g., making them more annoying in the case of the Game). This lends support to the idea that cross-modal approaches have value as a way of measuring the effects of application and usage contexts.

5 A CROSS-MODAL APPROACH VERSUS THE TARGET MODALITY

To gain insight into the differences that a full simulation of a target modality may have over a cross-modal approach for eliciting error characteristics, we conducted a follow-up study comparing the cross-modal approach from the first study with a full simulation of the mid-air gesture input in VR (pinch + directional gesture). The objective was to gain further insights into what aspects of the target modality may be missed in a cross-modal approach, or what may be of particular importance for understanding the acceptability of error with the target modality.

5.1 Experimental Protocol

The experimental protocol was consistent with the first study, with the following exceptions: (1) two experiential modalities were used, that is, a cross-modal version (or tablet) and the full VR simulation; and (2) the ordering of modality appearance was counter-balanced.

5.1.1 Apparatus. The cross-modal version was presented on a Samsung Galaxy Tab S6 Lite, with a 10.4 inch display size, running Android 11. The full VR simulation was presented on an Oculus Quest 2. To obtain recognition of the mid-air pinch + directional movement gesture, we leveraged the optical bare-hand tracking, stock on the Quest 2 devices. To allow for social/physical distancing, devices were delivered to participants and disinfected between use. The experimenter and participants conducted a video call to supervise the study and ensure proper functioning of the study software.

5.1.2 Participants and Procedure. Twelve participants (μ = 29.5 years, σ. = 8.72 years, range = 24 to 56 years) were recruited to participate in the study. Six participants identified as female, the remaining six as male. 11 participants were right handed, one participant was left handed. Participants were recruited through a university mailing list and word-of-mouth. Study length and compensation remained consistent with the first study.

A within-subjects design was followed with factors: Application (Endless Runner, Finger Painting), Modality (Cross-Modal, Full-Simulation), and Accuracy Scenario (P95/R95, P94/R80, P83/R95, P80/R80). Participants experienced each modality and application pairing counter-balanced using a Latin Square, and Accuracy Scenario randomized for each combination of (App:Modality). Following the study, participants were asked a series of open ended questions, designed to probe for what could and could not be understood from each modality.

5.2 Results: Likert Response Data

As in experiment 1, an align-and-rank transform (ART) [26, 44] factorial ANOVA was run on each of the five Likert scales (Acceptability, Usefulness, Effort, Frustration, and Overall UX).

- *Acceptability:* For acceptability, we found significant differences for Scenario ($F_{3,165} = 7.51, p < 0.001$) and Modality ($F_{1,165} = 6.18, p < 0.05$). Post-hoc tests using Tukey's HSD revealed significant differences between Scenarios (P95/R95 > P83/R95) and Scenarios (P95/R95 > P80/R80).
- *Usefulness:* For usefulness, significant differences were found for Scenario ($F_{3,165} = 5.15, p < 0.01$), Application ($F_{1,165} = 10.82, p < 0.01$), and Modality ($F_{1,165} = 10.68, p < 0.01$). Post-hoc tests using Tukey's HSD revealed a significant difference between Scenarios (P95/R95 > P83/R95) and Scenarios (P95/R95 > P80/R80).
- *Effort:* For effort, only Modality was significant ($F_{1,165} = 19.19, p < 0.001$).
- *Frustration:* Frustration revealed significance for both Scenario ($F_{3,165} = 5.58, p < 0.01$) and Modality ($F_{1,165} = 3.92, p < 0.05$. Post-hoc tests using Tukey's HSD revealed significant differences between Scenarios (P95/R95 < P83/R95) and Scenarios (P95/R95 < P80/R80), i.e., lower frustration for higher precision. Interaction effects were found for Application:Modality ($F_{1,165} = 0.85, p < 0.05$.
- *Overall UX:* For overall user experience, we found significant effects for Scenario ($F_{3,165} = 7.98, p < 0.001$) and Modality ($F_{1,165} = 12.65, p < 0.001$). Again, post-hoc tests using Tukey's HSD showed a significant difference between Scenarios (P95/R95 > P83/R95) and Scenarios (P95/R95 > P80/R80).

Raw Likert responses are depicted in Figure 6 using 100% stacked bar charts for each scenario per Likert question per App per Modality.

Recall that the goal of this second study was to contrast cross-modal error characteristic elicitation with error characteristic elicitation via the target modality. As a result, we wished to probe how various scenarios impacted Likert measures for the target modality in comparison to via cross-modal elicitation. To perform this analysis, we applied a numeric mapping to Likert responses (1 = strongly disagree; 7 = strongly agree). We then calculated the average values for each Likert measure for each modality for each application. This yielded 4 data points per modality (cross-modal or target modality) for each app (Paint and Game). We then performed a correlation analysis on the mean values per Scenario, Application, and Modality of each Likert scale using Pearson's correlation. Table 4 highlights these correlations. Results are mixed. We found significant correlations for Acceptability for the Game app, for Frustration for the Paint App, for Usefulness for the Game, and for Overall UX between measures taken for the cross-modal elicitation and the target modality elicitation.

5.3 Results: Qualitative Responses

The post-study questionnaire asked participants a series of questions designed to understand differences in experience between the two applications and modalities. Specifically, participants were asked four questions of the form: *Was there anything in the [VR, tablet] version that changed your understanding, or how you felt, about the*

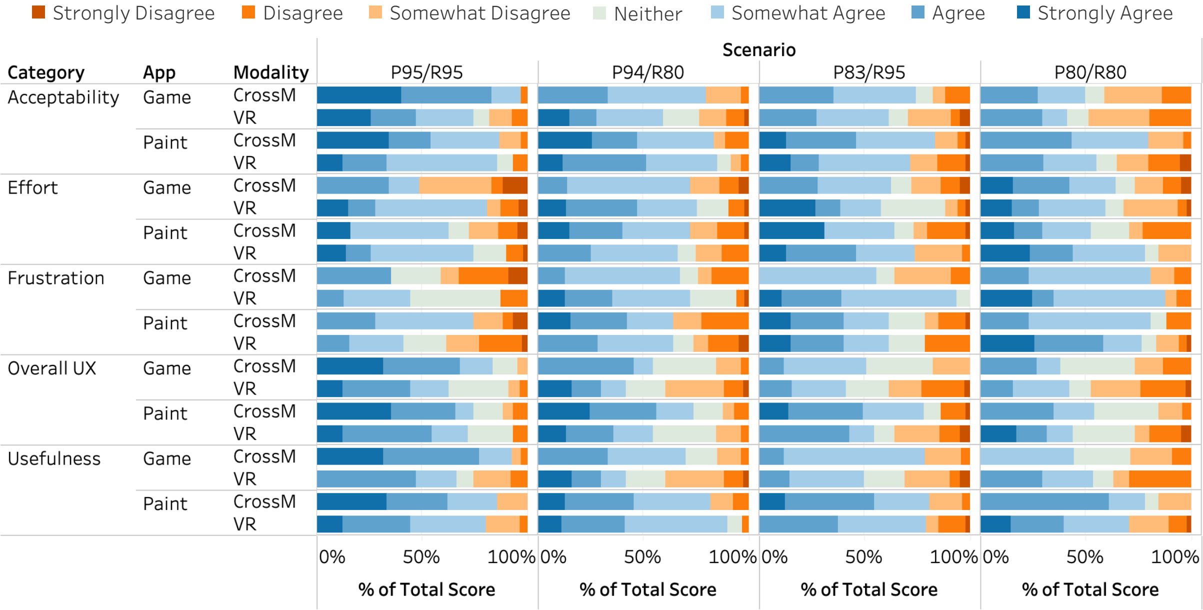

Fig. 6. Summary of Likert data across categories by Modality and Application.

Table 4. Results of correlation analysis of Likert measures for each app (Paint and Game) between modalities. Note that correlations are performed on only eight data points (four scenarios for each modality), so care must be taken in interpreting these values.

Likert Measure	Paint App	Game App
Acceptability	0.47	**1.00****
Effort	-0.50	0.39
Frustration	**1.00****	0.65
Usefulness	-0.14	**0.96***
Overall UX	**0.99***	**0.97***

experience and severity of [False Positive, False Negative] errors? Explain. They were also asked *Do you think the tablet version accurately communicated what the VR version would be like? If not, what was different or missing?*; *Did you feel anything was missing from the tablet version that would have helped you understand the VR version better? What was it?*; and *Can you think of any ways the tablet version could be changed to better communicate the VR interaction? What additions or changes might help with this?* Responses were analyzed for common themes that could suggest differences in the experience of input and errors between applications and modalities.

5.3.1 Apps across Modalities. As in the previous study, participants noted that errors were worse in the Game app than in the Paint app "due to the nature of failing the game" (P10). This held for both false positives "because [the game] was time-sensitive" (P7) and false negatives because it "meant that if I made a gesture and the application did not recognize it, I would fail" (P12).

When asked to contrast the tablet vs VR experience, two participants, P1 and P10 commented on the Paint app being more difficult in VR. P1 suggested that the lack of feedback when "touching" the canvas in the VR condition made the Paint app harder to use. This was apparently unrelated to the pinch+gestural input – as P1 noted that the gestures and errors experienced felt the "exact same" – but it is possible that the frustration that this created had a spillover effect and influenced participants' Likert responses regarding error acceptability.

5.3.2 VR vs. Tablet Gesture Input. Several comments suggested that the requirement of touching the tablet in the cross-modal condition changed the experience of providing input, as compared to the VR version. In response to the question regarding whether the tablet version accurately communicated what the VR version would be like, P10 replied: "because you could feel the tablet surface, it was a different experience than the VR version. It did help me get familiar with the motion of pinching (I did the tablet version before the VR version)" (P10).

Several participants raised the point of errors being more obvious in the tablet condition than VR. For some participants, this meant that they were less able to distinguish injected errors from user-caused errors in the VR condition (P5, P7, P10, P12). P7 and P12 suggested that this was due to their awareness of always being monitored in VR: "I sometimes wondered if I had moved just a little bit which had triggered a response" (P12).

However, mistakes were also possible on the tablet – space constraints resulted in participants accidentally triggering buttons, and feeling cramped performing gestures on the screen: "I kept on accidentally pressing buttons [on the tablet] but I didn't have that issue in VR. I found it harder to do the vertical-swipe gestures because there was less area on the screen to move in those directions [compared to horizontal in landscape view] and that limitation doesn't exist in VR" (P7).

The issue of comfort in each modality was raised by two other participants as well. P3 echoed the difficulty of the pinch + directional movement gesture on the tablet due to the friction on the surface of the tablet. In contrast, P9 commented that the tablet could not communicate fatigue experienced in VR, since "the hand has to be raised to be within sight of the VR headset, whereas your arm does not have to be raised to perform the gesture on the tablet" (P9).

Again, it is unclear how the above issues of comfort and fatigue may influence participants' assessment of errors, but to the extent that they might, it is interesting to consider how the cross-modal approach might be extended to take these factors into account. For instance, one could imagine a modified cross-modal approach that uses a much larger touch-screen surface mounted vertically in front of the user in a similar position to the field of view requirements that would be imposed by sensing the hands from the VR headset.

5.3.3 Expectations from Prior Modality Experience. Related to the points raised in the previous section, several comments suggested that participants' prior experiences with tablet and VR modalities influenced their experience of providing input in the study. P7 commented that they are used to the errors that come with hand tracking in VR, and so were more forgiving in the VR condition. Conversely, P12 cited their lack of experience in VR as leading them to question the source of false negative errors. This experience was also observed for the tablet variants, where participants found "false negatives were a lot harder in the painting app on the tablet to tolerate" (P8). It is unclear how much these expectations might influence participants' ratings of the acceptability of errors, but these comments do suggest it may be valuable to study the ability of expectation effects to influence responses given when trying to elicit acceptable error characteristics.

Expectations played out in other ways as well. When asked whether the tablet version accurately communicated what the tablet version would be like, P12 replied: "I think that [the tablet version] accurately communicated what the VR version would be like. Especially during the endless run game I found the VR versions more intuitive to pinch and release, compared to on a tablet version of that kind of game I would expect to just swipe one way or another (no pinching)" (P12).

The comment that the pinch gesture seemed less intuitive on the tablet version makes sense, given that the act of touching the tablet surface can play the role of a gesture delimiter on touch screens, so such gestures are not needed. This raises an interesting question – in a cross-modal approach, is it better to try and simulate the motor movements associated with the target modality as closely as possible (which would suggest recreating the pinching mechanic, even if it is awkward in the testing modality), or to choose an input that is functionally equivalent and intuitive in the testing modality (such as simple directional swipes)?

5.4 Study Synthesis

Synthesizing results from this study, consider, first, our Likert data. For measures of Acceptability, Usefulness, Frustration, and Overall UX, we see differences between error scenarios, and, more specifically, between the Scenarios (P95/R95 – P83/R95) and Scenarios (P95/R95 – P80/R80). In contrast, post-hoc tests did not reveal significant differences between Scenarios (P95/R95 – P94/R80) nor between Scenarios (P83/R95 – P80/R80). In other words, *precision* causes significant differences in measures of Acceptability, Usefulness, Frustration, and Overall UX, whereas there appears to be increased tolerance – based upon our statistical analysis – for differences in *recall*.

In practice, this is a useful result to inform system design. For example, techniques such as Katsuragawa et al.'s Bi-Level Thresholding [24] highlight how recall and precision can trade-off to produce both higher recognition rates (i.e. higher precision) and increased perceptions of reliability. In essence, we can tighten criterion functions or used repeated attempts – both of which harm recall – to increase precision and, based upon these results, to create higher ratings of Acceptability, Usefulness, and Overall UX while lowering Frustration. That said, differences do exist between touch and gesture, and our participants highlighted some of these differences, including the Paint app's increased difficulty in VR, the differences between surface and in-air interaction, and lower tolerance for false negatives (i.e. lower recall) in the tablet condition. We will return to these differences in our Discussion.

6 DISCUSSION

6.1 Understanding the Impact of Modality on Acceptable Error Characteristics

The motivation for the work described in this paper – as with the work of Kay et al. [25] – is centred around the challenge of eliciting acceptable error characteristics during early system design. Kay et al. leverage a text-based questionnaire to understand how and whether sensor-based error rates can be managed to inform decisions about engineering investment, whereas we explore eliciting error characteristics centred around recognition-based input techniques and explore a set of modalities (text, video, alternative input modalities, or the target modality) to elicit error characteristics.

To explore error elicitation for recognition-based input techniques, we present two studies. Our first study explores the difference between three modalities for error-elicitation: Kay's text-based questionnaire using scenarios [25], a video-based elicitation approach that depicts the dynamics of the interaction, and an alternative modality (e.g. touch) to assess reaction to different error characteristics in a target modality (mid-air gesture). Our second study contrasts the alternative or cross-modal approach with interaction in the target modality to probe both the benefits and risks of cross-modal elicitation.

At a high level, our results argue that:

- *Text descriptions, as in Kay et al., are insufficient on their own to elicit error characteristics for input techniques.* In our first study, alongside significant differences in Likert ratings, participants were much more comfortable with both video and cross-modal representations in terms of their ability to communicate aspects of the interaction specifically for recognition-based input techniques.
- *Cross-modal techniques can provide complementary information that can be useful to inform design.* In our second study, variations in precision resulted in statistically significant advantages in Acceptability,

Frustration, Usefulness, and Overall UX highlighting the importance of accuracy (as opposed to recall) when interpreting user actions.

Synthesizing the findings of the two studies, Table 5 shows error characteristics cited by participants as supported by the various modalities explored. The remainder of the discussion is structured around the findings highlighted in the table.

Table 5. Characteristics cited by participants for each modality.

Characteristic	Text	Video	Cross-Modal	Full Simulation
Communicates exact error rates	X			
Differences in accuracy conditions	X	X	X	X
Differences in applications/context		X	X	X
Interaction dynamics		X	X	X
Effort (physical and mental)			X	X
Emotional reaction to errors			X	X
Exact interaction mechanics				X

6.1.1 *Differences in Error Scenarios & Context.*

Consistent with Kay et al. [25], our study results demonstrate that participants provide different assessments of acceptability for different *accuracy scenarios* presented in the text modality. However, we also build upon findings Kay et al.

In our first study, our primary finding is that participants respond with different assessments of error characteristics for different modalities. In particular, in Table 5 we note that text communicates exact error rates (unlike the other modalities evaluated) and allows participants to note differences in precision and recall. This type of elicitation may work well in situations such as the smart alarm system highlighted in Kay et al.'s introduction that uses sensors to identify intruders and then needs to make a decision on notifying the homeowner via text or calling the police directly. In this situation, homeowners can, perhaps, use their imagination to understand how consequential different balances of precision and recall would be. In other words, they can evaluate the cost of false alarms versus the cost of missed intrusion to determine acceptable levels of precision and recall.

While potentially effective for a smart alarm system, our results in study one raise questions on the effectiveness of text for assessing the impact of different error characteristics for recognition-based input techniques. As one simple example, note that text-based elicitation struggles to discriminate differences across application or context. As another, our participants were more comfortable (11/12 and 12/12) in representing the impact of different error characteristics in both video and using cross-modal input than with text (7/12).

In our second study, we observed a difference in participants' assessments of precision versus recall in terms of impact on measures of Accuracy, Usefulness, Frustration, and Overall UX. The impact of lower recall was significantly less severe than lower precision. This may, in part, be because of participants' experiences with each modality (as noted in our qualitative results) and their experience with providing computer input in general. As a simple example, in a computer interface, if we click on an icon and nothing happens, almost subconsciously we immediately try again [23, 24]. Repeated recall failure can become a problem, but there is more tolerance; precision errors are significantly more severe [29], particularly when precision errors imply misrecognition and an incorrect action. This subtlety, an issue of interaction dynamics (see Table 5), was better captured by more interactive elicitations of error characteristics. Text seems to have a limited ability to support subtle distinctions in recall and precision.

Another area where text appears to struggle is in understanding the impact of errors in different applications or contexts. Here, video exhibits advantages because, with a visual depiction of the impact of errors during input, participants are better able to infer how variations in precision and recall may not symmetrically impact different applications. In contrast, text is less directly linked to the dynamics of interaction.

6.1.2 *Emotional Reaction to Errors.*

In participants' explanations of what distinguished the interactive modalities from text and video, we observed many qualitative comments related to emotional reactions, including terms such as "frustration", "calming", "fun", "enjoyable", and "annoying". This is also reflected in our Likert data, which suggests that frustration with particular error characteristics is best understood through interactive experiences, rather than non-interactive modalities such text descriptions or video demonstrations. This suggests a further value in interactive modalities for conveying how errors affect user frustration and other emotional reactions to errors, which is likely to also be important for assessing error acceptability.

6.1.3 *Exact Interaction Mechanics.*

The comments of several participants in both studies suggest that the differences in exact interaction mechanics created by the cross-modal approach led to some dissonance. Though participants were told that the intent of all the approaches was to simulate freehand mid-air pinch input in VR, one participant was surprised at errors occurring when they were not touching the tablet, and another commented that performing the pinch gesture on the tablet was more awkward than they imagined the full interaction would be. Our Likert data from the second study, and particularly our qualitative data, highlight this in additional detail. For example, participants noted that they were less tolerant of false negatives on the tablet because touching the screen is a deterministic act which should be detectable. In contrast, for the VR condition, participants noted an additional tolerance for system failure based on past experience with VR, and an ambiguity in attribution, i.e. a lack of certainty as to whether the error resulted from a failure of the system to interpret or an incorrect action of the user. Collectively, these are interesting examples of asymmetries that can arise in cross-modal elicitation of error characteristics.

6.2 Implications for Elicitation of Acceptable Error Characteristics

In Figure 2, we highlight a cost versus fidelity continuum, where text descriptions represent the lowest cost/lowest fidelity representation of error characteristics, and a full system simulation represents the highest fidelity/highest cost. We present video and cross-modal approaches as mid-points along this overall continuum. In this section, we argue that all of Text, Video, and Cross-Modal approaches can serve a valuable role in understanding acceptable error characteristics during system development, and provide guidance on how designers might use our findings.

Despite text's weaknesses in terms of discriminating applications or contexts, communicating interaction dynamics, and conveying effort and affective implications of errors, it could still be used at an early stage of design to elicit initial reactions from prospective users. However, we would argue that some initial feasibility work must be done prior to text-based elicitation, to understand the conditions under which errors may occur in a prospective application, their potential consequences, and a reasonable space of precision/recall values to test.

Video-based elicitation, like text, is also a very early stage technique for understanding error acceptability. Video, however, provides advantages because applications/contexts and interaction dynamics can be better conveyed, yielding more nuanced information. Participants noted this in our first study: almost half struggled with the fidelity of the text-based information as a representation of the final interaction, whereas video and cross-modal interactive approaches were – in the opinion of our participants – a more accurate modality to communicate representative aspects of interaction for understanding the effects of errors.

In addition, the results from Study 1 provide evidence that error characteristics elicited from our cross-modal simulation differ from error characteristics elicited from more passive observations by prospective users (e.g. reading text or watching video) and Study 2 supports this initial finding by highlighting correlations between cross-modal and target modality error scenarios.

Our findings also suggest that care must be taken in employing the cross-modal approach. In qualitative comments, our participants highlighted that they were less forgiving of errors in touch than in VR, based on their expectations of touch input. Furthermore, we found that factors that could influence experience of a system, such as effort and fatigue, can be missed in a cross-modal approach. When eliciting acceptable error characteristics via text, video, or a cross-modal approach, interaction designers should consider whether users' response to error characteristics are being driven by factors that are true to the target input modality, versus the testing representation. The spectrum of modalities discussed in this paper provides a range of approaches by which acceptable error characteristics might be elicited, which we hope will act as a starting point for developing rigorous approaches to address this challenge.

The focus of this paper has been on investigating modalities that could be used to elicit acceptable error characteristics for input techniques, but an equally important question is how to design studies for this purpose. As in Kay et al.'s work, a method could be used in which prospective users are presented with a range of accuracy scenarios for an application, and their responses are synthesized to produce an acceptability of accuracy metric [25]. Given that many of the modalities studied in this paper have a higher cost to test than the text modality used by Kay et al., evaluations for input techniques might require some initial work to narrow the range of accuracy scenarios. In the case where a designer is studying an input technique under active development, the accuracy characteristics of current prototypes could be used to guide the range of accuracy scenarios to test.

6.3 Limitations and Future Work

This work has several limitations, which should be addressed in future work. The most significant limitation is that, since we only tested one input technique, it remains to be seen whether our findings will generalize more broadly. While the tested input technique (pinch + directional gesture) mapped consistently to a touch screen cross-modal implementation, it may be more challenging to find a mapping for other novel input techniques, such as interactive textiles, or complex gestures based on body pose. More broadly, it is interesting to consider what a cross-modal approach might look like when applied to recognition-based systems that use non-physical input, such as speech or gaze-based input.

Another challenge is that the error characteristics of many input techniques cannot be captured with precision/recall metrics alone. For example, input techniques may rely on multi-class classification, in which the recognizer attempts to determine which of a *set* of gestures a user has performed, rather than whether or not a single gesture has been performed. In this situation, the system can make additional types of errors – such as mistaking one gesture for another, similar to how a speech recognizer may mistake an utterance of one word or phrase for another. The modalities explored in the present work may still be applicable in these cases, but new approaches may be required to expose participants to the full range of errors that the system could make, and to define and elicit acceptable error characteristics in response.

Finally, though we would anticipate that the "closer" the cross-modal approach is to the target modality, the better (or more accurate) the representation of acceptable error characteristics will be, this remains an open research question – for instance, could a lower fidelity prototyping tool using simpler input (such as a button press) be as useful in eliciting such information? Future work is needed to explore this possibility, and to more broadly investigate the various types of errors that can occur with input techniques, and the factors that influence how negatively these errors will influence a user's experience of an interactive system.

7 CONCLUSION

To enable designers and researchers to understand the acceptable error characteristics of input techniques early in the design process, this paper has contributed a design space of potential approaches on a continuum of optimizing for cost vs. fidelity, and two studies to investigate the benefits and drawbacks of four modalities across this space. For simulating a mid-air virtual reality input technique, several benefits were found for modalities which allow users to interactively try out a representation of a technique, as compared to non-interactive modalities. These findings lay the groundwork for understanding accuracy requirements early in the development of novel input techniques, to guide the development and improvement of sensing and recognition approaches.

REFERENCES

[1] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: Enhancing Movement Training with an Augmented Reality Mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 311–320. https://doi.org/10.1145/2501988.2502045

[2] Ravin Balakrishnan and I. Scott MacKenzie. 1997. Performance Differences in the Fingers, Wrist, and Forearm in Computer Input Control. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 303–310. https://doi.org/10.1145/258549.258764

[3] Olivier Bau and Wendy E. Mackay. 2008. OctoPocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology* (Monterey, CA, USA, 2008-10-19) *(UIST '08)*. Association for Computing Machinery, 37–46. https://doi.org/10.1145/1449715.1449724

[4] Michel Beaudouin-Lafon and Wendy E. Mackay. 2012. Prototyping Tools and Techniques. In *The Human Computer Interaction Handbook* (3 ed.), Julie A. Jacko (Ed.). CRC Press, Boca Raton, FL, Chapter 47, 1081–1104. https://doi.org/10.1201/b11963-55

[5] Marion Buchenau and Jane Fulton Suri. [n.d.]. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques* (New York City, New York, USA, 2000-08-01) *(DIS '00)*. Association for Computing Machinery, 424–433. https://doi.org/10.1145/347642.347802

[6] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '93)*. Association for Computing Machinery, New York, NY, USA, 193–200. https://doi.org/10.1145/169891.169968

[7] Fred Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (09 1989), 319–. https://doi.org/10.2307/249008

[8] William Delamare, Thomas Janssoone, Céline Coutrix, and Laurence Nigay. 2016. Designing 3D Gesture Guidance: Visual Feedback and Feedforward Design Options. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Bari, Italy) *(AVI '16)*. Association for Computing Machinery, New York, NY, USA, 152–159. https://doi.org/10.1145/2909132.2909260

[9] P. Dourish. 2004. *Where the Action Is: The Foundations of Embodied Interaction.* MIT Press. https://books.google.ca/books?id=-TRWc0PA9e4C

[10] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. [n.d.]. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017) *(CHI '17)*. ACM, 278–288. https://doi.org/10.1145/3025453.3025739

[11] Pelle Ehn and Morten Kyng. 1992. Cardboard Computers: Mocking-It-up or Hands-on the Future. In *Design at Work: Cooperative Design of Computer Systems*. L. Erlbaum Associates Inc., USA, 169–196.

[12] World Leaders in Research-Based User Experience. 1994. *Guerrilla HCI: Article by Jakob Nielsen.* https://www.nngroup.com/articles/guerrilla-hci/ Library Catalog: www.nngroup.com.

[13] World Leaders in Research-Based User Experience. 2009. *Discount Usability: 20 Years.* https://www.nngroup.com/articles/discount-usability-20-years/ Library Catalog: www.nngroup.com.

[14] Dustin Freeman, Hrvoje Benko, Meredith Ringel Morris, and Daniel Wigdor. [n.d.]. ShadowGuides: visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Banff, Alberta, Canada, 2009-11-23) *(ITS '09)*. Association for Computing Machinery, 165–172. https://doi.org/10.1145/1731903.1731935

[15] Dustin Freeman, Hrvoje Benko, Meredith Ringel Morris, and Daniel Wigdor. 2009. ShadowGuides: Visualizations for in-Situ Learning of Multi-Touch and Whole-Hand Gestures. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (Banff, Alberta, Canada) *(ITS '09)*. Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/1731903.1731935

[16] Saul Greenberg and Chester Fitchett. [n.d.]. Phidgets: easy development of physical interfaces through physical widgets. In *Proceedings of the 14th annual ACM symposium on User interface software and technology* (Orlando, Florida, 2001-11-11) *(UIST '01)*. Association for Computing Machinery, 209–218. https://doi.org/10.1145/502348.502388

[17] Carl Gutwin, Andy Cockburn, and Benjamin Lafreniere. 2015. *Testing the Rehearsal Hypothesis with Two FastTap Interfaces.* 223–231.

[18] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

[19] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring Sensor-Based Interactions by Demonstration with Direct Manipulation and Pattern Recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1240624.1240646

[20] Jay Henderson, Sachi Mizobuchi, Wei Li, and Edward Lank. 2019. Exploring Cross-Modal Training via Touch to Learn a Mid-Air Marking Menu Gesture Set. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 8, 9 pages. https://doi.org/10.1145/3338286.3340119

[21] Ankit Kamal, Yang Li, and Edward Lank. [n.d.]. Teaching motion gestures via recognizer feedback. In *Proceedings of the 19th international conference on Intelligent User Interfaces* (Haifa, Israel, 2014-02-24) *(IUI '14)*. Association for Computing Machinery, 73–82. https://doi.org/10.1145/2557500.2557521

[22] Ankit Kamal, Yang Li, and Edward Lank. 2014. Teaching Motion Gestures via Recognizer Feedback. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. Association for Computing Machinery, New York, NY, USA, 73–82. https://doi.org/10.1145/2557500.2557521

[23] Keiko Katsuragawa, Ankit Kamal, and Edward Lank. 2017. Effect of motion-gesture recognizer error pattern on user workload and behavior. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. https://doi.org/10.1145/3025171.3025234

[24] Keiko Katsuragawa, Ankit Kamal, Qi Feng Liu, Matei Negulescu, and Edward Lank. 2019. Bi-Level Thresholding: Analyzing the Effect of Repeated Errors in Gesture Input. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 15 (April 2019), 30 pages. https://doi.org/10.1145/3181672

[25] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2702123.2702603

[26] Matthew Kay and Jacob O. Wobbrock. 2020. Aligned Rank Transform for Nonparametric Factorial ANOVAs. https://doi.org/10.5281/zenodo.594511

[27] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300641

[28] Matthias Kranz, Paul Holleis, and Albrecht Schmidt. 2010. Embedded Interaction: Interacting with the Internet of Things. *IEEE Internet Computing* 14, 2 (2010), 46–53. https://doi.org/10.1109/MIC.2009.141

[29] Ben Lafreniere, Tanya R. Jonker, Stephanie Santosa, Mark Parent, Michael Glueck, Tovi Grossman, Hrvoje Benko, and Daniel Wigdor. 2021. False Positives vs. False Negatives: The Effects of Recovery Time and Cognitive Costs on Input Error Preference. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 54–68. https://doi.org/10.1145/3472749.3474735

[30] David Ledo, Fraser Anderson, Ryan Schmidt, Lora Oehlberg, Saul Greenberg, and Tovi Grossman. [n.d.]. Pineal: Bringing Passive Objects to Life with Embedded Mobile Devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA, 2017-05-02) *(CHI '17)*. Association for Computing Machinery, 2583–2593. https://doi.org/10.1145/3025453.3025652

[31] David Ledo, Jo Vermeulen, Sheelagh Carpendale, Saul Greenberg, Lora Oehlberg, and Sebastian Boring. [n.d.]. Astral: Prototyping Mobile and Smart Object Interactive Behaviours Using Familiar Applications. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA, 2019-06-18) *(DIS '19)*. Association for Computing Machinery, 711–724. https://doi.org/10.1145/3322276.3322329

[32] Yang Li. [n.d.]. Protractor: a fast and accurate gesture recognizer. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA, 2010-04-10) *(CHI '10)*. Association for Computing Machinery, 2169–2172. https://doi.org/10.1145/1753326.1753654

[33] Microsoft. 2020. Gestures - Mixed Reality Toolkit Documentation. https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/Input/Gestures.html

[34] William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K. Dey, and Min Kyung Lee. [n.d.]. A fieldwork of the future with user enactments. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom, 2012-06-11) *(DIS '12)*. Association for Computing Machinery, 338–347. https://doi.org/10.1145/2317956.2318008

[35] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation accuracy is good, but high controllability may be better. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300750

[36] Dean Rubine. 1991. Specifying Gestures by Example. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '91)*. Association for Computing Machinery, New York, NY, USA, 329–337. https://doi.org/10.1145/122718.122753

[37] Valkyrie Savage, Colin Chang, and Björn Hartmann. 2013. Sauron: Embedded Single-Camera Sensing of Printed Physical User Interfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 447–456. https://doi.org/10.1145/2501988.2501992

[38] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: Projected Visualizations for Hand Movement Guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/2207676.2207702

[39] Seiichi Uchida and Hiroaki Sakoe. 2005. A survey of elastic matching techniques for handwritten character recognition. *IEICE transactions on information and systems* 88, 8 (2005), 1781–1790.

[40] Viswanath Venkatesh and Fred Davis. 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46 (02 2000), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

[41] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

[42] Daniel Wigdor and Dennis Wixon. 2011. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[43] Andrew D. Wilson. 2006. Robust Computer Vision-Based Detection of Pinching for One and Two-Handed Gesture Input. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland) *(UIST '06)*. Association for Computing Machinery, New York, NY, USA, 255–258. https://doi.org/10.1145/1166253.1166292

[44] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[45] J. Zimmerman. 2005. Video sketches: exploring pervasive computing interaction designs. *IEEE Pervasive Computing* 4, 4 (2005), 91–94.