False Positives vs. False Negatives

The Effects of Recovery Time and Cognitive Costs on Input Error Preference

Ben Lafreniere

Facebook Reality Labs Research, Toronto, ON, Canada, benlafreniere@fb.com

Tanya R. Jonker

Facebook Reality Labs Research, Seattle, WA, USA, tanya.jonker@fb.com

Stephanie Santosa

Facebook Reality Labs Research, Toronto, ON, Canada, ssantosa@fb.com

Mark Parent

Facebook Reality Labs Research, Toronto, ON, Canada, mrkprnt@fb.com

Michael Glueck

Facebook Reality Labs Research, Toronto, ON, Canada, mglueck@fb.com

Tovi Grossman

University of Toronto, Toronto, ON, Canada, tovi@dgp.toronto.edu

Hrvoje Benko

Facebook Reality Labs Research, Seattle, WA, USA, benko@fb.com

Daniel Wigdor

Facebook Reality Labs Research and University of Toronto, Toronto, ON, Canada, dwigdor@fb.com

Existing approaches to trading off false positive versus false negative errors in input recognition are based on imprecise ideas of how these errors affect user experience that are unlikely to hold for all situations. To inform dynamic approaches to setting such a tradeoff, two user studies were conducted on how relative preference for false positive versus false negative errors is influenced by differences in the temporal cost of error recovery, and high-level task factors (time pressure, multi-tasking). Participants completed a tile selection task in which false positive and false negative errors were injected at a fixed rate, and the temporal cost to recover from each of the two types of error was varied, and then indicated a preference for one error type or the other, and a frustration rating for the task. Responses indicate that the temporal costs of error recovery can drive both frustration and relative error type preference, and that participants exhibit a bias against false positive errors, equivalent to ~1.5 seconds or more of added temporal recovery time. Several explanations for this bias were revealed, including that false positive errors impose a greater attentional demand on the user, and that recovering from false positive errors imposes a task switching cost.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI; Interaction techniques.

UIST '21, October 10–14, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8635-7/21/10…$15.00

https://doi.org/10.1145/3472749.3474735

Additional Keywords and Phrases: Error perception; gesture recognition; recognizer thresholds; probabilistic input; utility models; input ambiguity

ACM Reference Format:

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 UIST '21: ACM Symposium on User Interface Software and Technology, October 10–13, 2021, Virtual Event. ACM, New York, NY, USA.

1 INTRODUCTION

Many novel and emerging input techniques have the potential to misinterpret the user's intentions, due to limitations in sensing hardware (e.g., occlusion issues in computer vision), imperfect recognition (e.g., due to models trained on limited data), or ambiguities in the input (e.g., the "fat finger" problem on touchscreens). A particular challenge with input techniques that use a continuous data stream is correctly discriminating intentional input actions from all other user behavior. When this fails, two types of errors can occur – *false positives*, where the system recognizes an input action when the user did not intentionally perform one, and *false negatives*, where the system fails to recognize an input action that was intentionally performed by the user.

In gesture recognition, the most common approach for discriminating intentional input is to set a threshold on the score output by the recognizer – scores above the threshold trigger the action mapped to the gesture, while scores below it do not. However, false positive and false negative errors can still occur if the threshold does not perfectly separate intentional input actions from other user behavior. Approaches to limiting these errors include choosing gestures that are unlikely to naturally occur [7, 23], adding delimiter gestures [8, 21], and using a *bi-level thresholding* approach, in which a restrictive threshold is used for initial attempts at performing a gesture, to cut down on false positive errors, coupled with a more permissive threshold following 'near misses', to enable the user to succeed when trying the gesture a second time [10, 16]. While these approaches can reduce the occurrence of errors, they do not take into consideration the actions that happen (or don't happen) as a result. Just as a mouse click can activate many different actions depending on the cursor's location, a single gesture could activate many different actions depending on context, and the cost of error recovery will depend on this gesture-action mapping. This is important because, while false positive errors are often considered worse than false negative errors [10], it is easy to imagine situations where a false negative error has a high cost (e.g., the user is playing a fast-paced game where every second counts; or trying to beat an auto-play timeout on a streaming site), or a false positive error has a low cost (e.g., when the accidental action is easily reversed).

Motivated by the above, the present work investigates (1) how the temporal (i.e., time) cost of recovering from false positive and false negative errors influences error type preference; (2) whether there are hidden cognitive costs associated with these error types when temporal cost is accounted for; and (3) how the higher-level task the user is performing influences these costs. The idea is to lay the foundation for error-cost aware input recognizers, capable of dynamically adjusting their thresholds to prevent costly errors and optimize the user experience.

An experimental task was developed in which participants search a grid of tiles and select items using the mouse. False positive and false negative errors are injected at a controlled rate, and the temporal cost of recovery for each of the two error types is manipulated. By eliciting error type preferences over a range of differences in temporal costs, we can model the effects of temporal cost on error type preference *and* reveal hidden cognitive costs and biases not explained by temporal cost alone. A first study was run with this standard task, followed by a second study with two additional task variants – one with added time pressure, and one with added attentional demands.

The results of these studies revealed several novel findings on input errors and error type preference:

- The temporal costs of error recovery can drive both relative error type preference and frustration.
- When temporal costs of recovery are equivalent, users exhibit a bias against false positive errors, which can be equivalent to 1.5 seconds or more of added temporal cost, suggesting that the hidden cognitive costs of false positive errors are greater than those of false negative errors.
- The bias against false positive errors is in-part driven by the attentional demands of noticing when false positive errors have occurred.
- Clusters of error occurrences (i.e., "peak effects") and the error type experienced at the end of a block (i.e., "end effects") can influence error type preferences provided retrospectively after an experience.

Collectively, these results have implications for the design of gesture recognizers, the user interfaces for recognition-based input systems, and research methods for understanding the effects of recognizer errors.

2 RELATED WORK

This work complements and extends existing work on recognition metrics and thresholds, as well as work developing utility models and investigating cognitive biases in human-computer interaction.

2.1 Recognition Metrics and Thresholds

The overall goal of gesture recognition algorithms is to support high rates of both *precision* and *recall* [15, 18, 26]. In this context *precision* is the percentage of reports by the recognizer that a gesture has occurred that are correct, whereas *recall* is the percentage of performances of the gesture that are successfully caught by the recognizer. A lower precision means more false positives (FPs), while a lower recall means more false negatives (FNs). Precision and recall are sometimes combined into an F_1-score (the harmonic mean of precision and recall), or alternative F-measures which put greater weight on precision or recall [27].

The danger in focusing too closely on precision, recall, and overall error rates is that FP and FN errors may have very different consequences for user experience. In the information retrieval community it is well-established that, for many applications, precision and recall are not equally important to the user [22]. Recent work in the ubiquitous computing literature has proposed a method to elicit a weighting of precision and recall at design time, which can guide the development of classifier algorithms [11]. Dove et al. called for work to understand how precision and recall impact UX in systems with machine learning [5], and Kocielnik et al. [12] demonstrated the importance of finding the right balance between FP and FN errors in AI-based systems, and investigated design techniques to set user expectations and mitigate the impact of such errors.

Beyond balancing precision and recall at the application level, it may be valuable to change the balance in real-time. Negulescu et al.'s bi-level thresholding approach to gesture recognition biases toward higher precision most of the time (i.e., fewer FPs at the cost of more FNs), but relaxes the criterion for recognition after "near misses" in which the recognizer score comes close to the threshold, to enable users to succeed on a second attempt of a gesture following a FN [10, 16]. Katsuragawa et al. demonstrate that this approach can enhance precision and recall, and also improves user experience by reducing instances where users encounter the same error more than once in quick succession [10].

While past work has highlighted the importance of tuning the precision-recall tradeoff based on how FP and FN errors impact user experience, the goal has typically been a single tuning per application. Bi-level thresholding accounts for some context but does not consider the recovery cost of FP and FN errors to be dynamic. In contrast, the present work lays the groundwork for a precise and generalizable model of how error preference changes with the temporal cost of recovery, to inform dynamic approaches to setting recognizer thresholds.

2.2 Utility Models and Cognitive Biases

Horvitz's foundational work on mixed-initiative interfaces [9] proposed that an intelligent agent's decisions regarding action versus inaction should be based on expected utility, taking into account the cost of misinterpreting the user's goals. More recently, Banovic et al. used expected utility theory to model how the cost of error influences pointing behavior [1], and Quinn demonstrated that a number of cognitive biases apply to users during human-computer interactions, and that their effects can be captured in economic models of utility [17].

Subjective factors have also been shown to be important in how users respond to the accuracy and error characteristics of interactive systems. Roy et al. demonstrated that users will tolerate lower accuracy if they are afforded greater controllability in systems with intelligent assistance [20]. The intelligibility and interpretability of software that uses machine learning has been identified as key factors to user tolerance of errors [5, 24]. More broadly, studies of 'peak-end' effects have demonstrated that retrospective assessments of experience can be influenced by the sequence in which otherwise identical events occur [4, 6]. Though not the main focus of the present work, the studies in this paper also find evidence of peak-end effects.

In summary, past work has identified several subjective factors that can influence the perception and acceptability of errors, but there has yet to be an investigation of the relationship between temporal costs and user preference for FP versus FN errors, or whether there are inherent differences in how these error types impact user experience. This paper reports the first study to capture this relationship, and also provides further insights into hidden costs and biases in how these errors are experienced by users.

3 STUDY SYSTEM

This paper focuses on a scenario where a single command gesture can be used to activate different actions, as might occur in an AR/VR system where a single free-hand gesture could invoke many commands, depending on the hand's position in the virtual environment. This also mirrors standard WIMP interactions, where the cursor location provides context, and the mouse click acts as the command "gesture". A mouse-based interaction was used for the studies reported in this paper because, as a highly reliable input method, FP and FN errors could be injected at a tightly controlled rate.

A challenge to systematically studying FP and FN errors is that, typically, the conditions in which these error types occur are different, as are the actions required to recover from them. In an FN error the user has intentionally performed an action to provide input to the system, and the system has mistakenly ignored that action. Recovery involves retrying the input action, along with any setup needed to get back to a state where this is possible. In contrast, with an FP error, the user did not intentionally provide input to the system, but the system acts as though they did. Recovery involves noticing that the error has occurred and reversing any undesirable consequences of the unintended action. To investigate the effects of these errors, a study task was developed in which recovery from FN and FP errors require an identical sequence of actions, and the temporal cost associated with each error type can be manipulated.

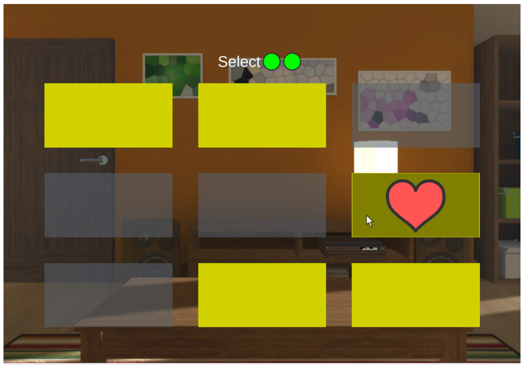


Figure 1. The study task interface.

The study task (Figure 1) involved finding and selecting tiles containing target items using the mouse. On each "page", five randomly selected tiles in a 3×3 grid are "enabled" (indicated in yellow). The user is instructed to select a specified number of target items (e.g., "*Select 2 green circles*"). The user can reveal the contents of an enabled tile by dwelling the cursor on it for 1.25 seconds; during this "hover delay", a radial progress indicator fills, and then the tile flips over to reveal one of six icons (a green circle, red heart, orange triangle, yellow star, blue moon, or purple plus). Once revealed, the user has a short time (1.25 seconds) to select the tile with a mouse click, after which the tile closes (flips back over), regardless of whether the item has been selected or not. Selected tiles are indicated with a blue outline. When the specified number of target items has been selected, the system proceeds to the next page, with a new random selection of five enabled tiles and a new number of target items to select.

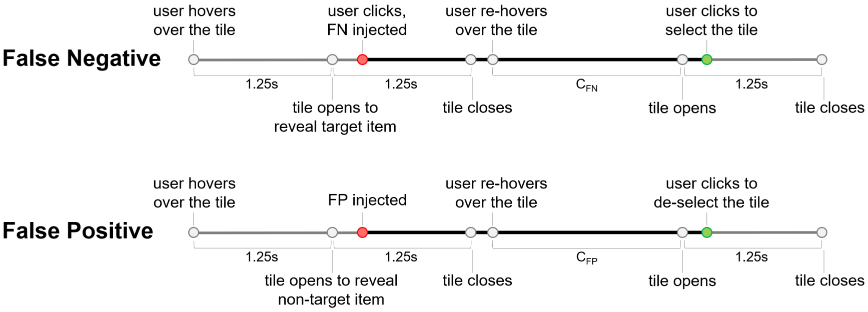

False Negative
user hovers over the tile
user clicks, FN injected
user re-hovers over the tile
user clicks to select the tile
1.25s
1.25s
C_{FN}
1.25s
tile opens to reveal target item
tile closes
tile opens
tile closes
False Positive
user hovers over the tile
FP injected
user re-hovers over the tile
user clicks to de-select the tile
1.25s
1.25s
C_{FP}
1.25s
tile opens to reveal non-target item
tile closes
tile opens
tile closes

Figure 2. Timelines from error injection (red) to recovery (green) for each error type.

FN errors can be injected by the system when the user has opened a tile containing the target item and clicked to select it (Figure 2 top). The system acts as though no click was performed, preventing the user from selecting the item before the tile closes. To retry, the user must first re-open the tile. To manipulate the temporal cost of error recovery, the typical hover delay for opening the tile is replaced with a delay of duration C_{FN}. To ensure that FN errors are dealt with immediately, the system displays a message on the tile, "*Target icon not selected. Select it to continue.*" and requires the user to correct it before proceeding to check other tiles.

FP errors can be injected by the system when the user has revealed a non-target item (Figure 2 bottom). Before the user can move the cursor off the tile, the system acts as though a click was performed, selecting the item and providing the standard click feedback (discussed below). To de-select the item, the user must first re-open the tile. To manipulate the time-based cost of recovery, the typical hover delay to open the tile is replaced with a delay of duration C_{FP}. To ensure that FP errors are dealt with immediately, the system displays a message, "*Non-target item selected. Deselect it to continue.*" and requires the user to correct it before proceeding to check other tiles.

In the above approach, the actions to recover from the two error types are consistent, and the study system can independently manipulate the error rate and temporal cost of recovery for each of the error types. Moreover, it does so while preserving the character of FN and FP errors – FN errors still occur in response to an intentional action to select a tile, whereas FP errors occur when such a selection action is not intended.

3.1 Error Injection Approach

Each participant experienced a number of "blocks" of the above task, consisting of 40 tile openings over a number of pages. Precisely controlling the number of FP and FN errors in each block was important for consistency. To achieve this, the sequence of items to be revealed when opening successive tiles was pre-determined at the start of each block, such that across all pages 50% of tiles would reveal a target item *regardless of the order in which the user opened the tiles.* For the tiles that revealed target items, a random sequence of true/false values was generated, specifying whether an FN error would be injected when the user attempted a selection. The sequence contained the correct number of 'true' and 'false' values to create the intended number of FN errors, in a randomized order. A similar procedure was used to generate a sequence of FP errors for tiles that revealed non-target items.

3.2 Input Modality and Feedback

Mouse input was selected for several reasons. First, a properly working mouse is free of both FN and FP errors, which enabled the rate of these errors to be entirely controlled through error-injection. Second, mouse input is highly accurate for targeting, reducing user errors. Finally, it enabled the studies to be deployed remotely, which provided efficient and safe data gathering during the global COVID-19 pandemic.

A challenge of using mouse input is that the main feedback mechanism for click input is the mechanical feedback from the button on the mouse device. This posed a challenge for injecting FP errors, which cannot provide such feedback. To address this, when a click occurs (either by the user clicking, or an injected FP error), a white circle is displayed around the click location, to indicate that the system has received click input.

To prevent rapid clicking, which could enable participants to recover from FP and FN errors in the study task without going through the process of re-opening a tile described above, a 1.25 second lockout is imposed after legitimate clicks, FP clicks, and FN errors. During this time, the cursor outline is temporarily changed from black to grey to communicate that the cursor is in the lockout state.

4 STUDY 1

Using the study system just described, an experiment was conducted to understand the relative costs of FP vs. FN errors, and gain insights into hidden costs and drivers of the user experience of these errors. The study was run on the Amazon Mechanical Turk platform, which provides access to a large and diverse participant pool [19].

4.1 Participants

52 participants were recruited and compensated $12 USD for the experiment, which took 60-75 minutes to complete. After filtering out participants that did not complete the study, and participants whose duration for the main task

conditions was greater than 3 standard deviations above the mean, or whose number of user errors was greater than 3 IQRs above the median, 44 participants were included in the analysis (Table 1).

Table 1. Demographics for Study 1 (N=44)

Age	M = 38 (SD 10, range: 21 to 61)
Gender	30 Male, 14 Female
Handedness	38 Right, 6 Left
Pointing Hand	41 Right, 3 Left
Computer Hardware	24 Laptop, 20 Desktop
Pointing Device	44 Mouse
Gaming Frequency	Daily (9), Weekly (16), Monthly (10), Yearly or less (9), Never (0)

4.2 Study Design and Procedure

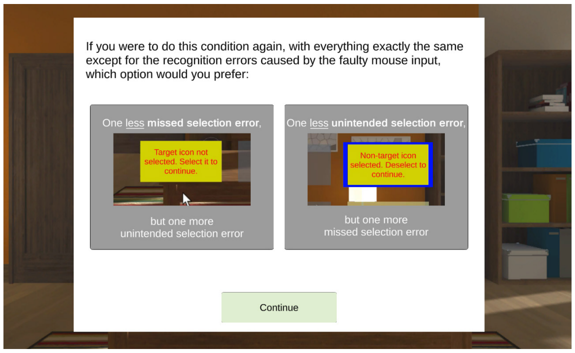
The experiment followed a within-subjects design with factor COST REGIMEN. Each cost regimen comprised a (C_{FN}, C_{FP}) pair defining the delay in seconds to reopen a tile when recovering from FN and FP errors, respectively. In choosing the reopening costs to test, we wanted (1) costs that were realistic to the time it takes to recover from common FP errors, such as accidentally opening a dialog, with recovery taking <5 seconds; and (2) a consistently spaced set of costs with a "resolution" fine enough to detect subtle biases against an error type, but coarse enough that users would be able to perceive differences between conditions. Pilot testing revealed that the range of 0.25s to 3.25s in 0.5s increments met these requirements. The cost regimens for the study are indicated in Table 2 – these values were generated by fixing one cost at 1.75 seconds (the midpoint of the range) and varying the other cost across the full range. This set of cost regimens results in seven unique deltas between C_{FP} and C_{FN} (-1.5 to 1.5 in 0.5 second increments).

Table 2. Cost regimens for Study 1.

C_{FN} \ C_{FP}	0.25	0.75	1.25	1.75	2.25	2.75	3.25
0.25				✓			
0.75				✓			
1.25				✓			
1.75	✓	✓	✓	✓	✓	✓	✓
2.25				✓			
2.75				✓			
3.25				✓			

For each condition, participants completed the tile selection task, opening 40 tiles over a number of pages. As mentioned previously, the study system was designed such that 50% of tiles would reveal a target item, and 50% a non-target item. In 6 target reveals, an FN error would be injected when the participant tried to click the item; and in 6 non-target reveals an FP would be injected. Errors were not injected during error recovery actions – while it would be realistic to do so, not injecting errors during recovery actions allowed for tighter control over the temporal cost of recovery, and greater consistency across blocks.

After each condition, the participant was asked three questions, which form the dependent measures for the study. First, they were asked "*If you had to do this condition again, with everything exactly the same except for the recognition errors caused by the faulty mouse input, which option would you prefer: one less missed selection error, but one more unintended selection error; or one less unintended selection error, but one more missed selection error*" (Figure 3). Next, they were asked to rate the strength of their preference (5-point scale with levels "[No, Weak, Moderate, Strong, Very strong] preference"). Finally, they were asked to rate their frustration level during the block (7-point scale, from "Very Low" to "Very High").


If you were to do this condition again, with everything exactly the same except for the recognition errors caused by the faulty mouse input, which option would you prefer:

One less **missed selection error**,
Target icon not selected. Select it to continue.
but one more unintended selection error
One less **unintended selection error**,
Non-target icon selected. Deselect to continue.
but one more missed selection error
Continue

Figure 3. The dialog used to elicit error type preference after each condition. The thumbnail on each button was a short 5-second animation demonstrating the error type.

In terms of the overall study procedure, participants started by completing a demographics questionnaire and receiving instructions for the study task. The instructions included text and short video clips to explain: the user's goal in the task, the mechanics for opening tiles and selecting items, the lock-out delay after a click, the two error types the user would encounter, and how to recover from each. This was followed by five short practice blocks: one with no injected errors, one with 100% FN errors, one with 100% FP errors, and two blocks demonstrating how the time to reopen tiles after each of the errors can vary, with cost regimens (0.5, 3.0) and (3.0, 0.5) respectively. Next, participants started the main data gathering portion of the experiment, in which they completed a block for each of the 13 cost regimen conditions (order counterbalanced across participants in a balanced Latin square). Each condition was followed by a break of at least 10 seconds. A post-study questionnaire asked which type of error the participant found more frustrating overall, and what they disliked about each of the error types.

4.3 Data Analysis Approach

In terms of the general data analysis approach, error preferences were analyzed based on how participants' error preference responses (i.e., which error-type they preferred, and their strength rating for that preference) varied in response to *Reopen_Delta* – the difference in the manipulated tile reopening time for FP and FN errors (i.e., $C_{FP} - C_{FN}$). Reopen_Delta captures how much greater the time cost of an FP error is as compared to an FN error in a block. If participants' only consideration was time cost, they would prefer FP errors when Reopen_Delta is less than zero, and FN errors when it is greater than zero. Deviations from this behavior reveal a bias not explained by time cost alone. Specifically, the bias can be estimated by modeling how error preference responds to Reopen_Delta and then examining the *indifference point* – the value of Reopen_Delta at which error preference shifts from FP to FN. Similar methods have been used in past work to quantify cognitive biases associated with interaction techniques [17].

Data analysis was performed using mixed-effects models, as our measurements were repeated within participants. Each analysis included a random intercept per participant; it did not include a random slope due to limited samples. Kenward-Rogers approximation was used to compute degrees of freedom for F-tests.

Initial analyses revealed that some participants confused the two error types. To address this, any participants who reported preference for FP errors in the two blocks with the highest C_{FP} relative to C_{FN} *and* preference for FN errors in the two blocks with the highest C_{FN} relative to C_{FP} were removed prior to analyses (1/44 participants). Blocks for which the target number of errors (6 FP, 6 FN) was not injected were also removed (this could happen if a participant made many user errors, but was rare, occurring in only 1/572 blocks).

4.4 Results

To provide a general sense of the tasks experienced by participants, and to validate the effectiveness of the error injection approach and temporal cost manipulation, this section starts by presenting task time, error rates, and a recovery time for the two error types. This is followed by a regression analysis of error type preferences and frustration ratings, and an analysis of post-study questionnaire responses.

4.4.1 Task Time, Error Injection, and Recovery Time

The mean duration of a block (not including practice blocks) was 183 seconds (SD 53, min 128, max 910). The high max value may be explained by participants “multi-tasking” or taking breaks during blocks, which is difficult to control in remote studies. In terms of injected errors, the correct number of errors (6 each of FP and FN) was injected in 571/572 blocks (99.8%). User-caused errors were rare – across all blocks, users had a mean of 5.3 (SD 6.9) instances where they failed to click a target item before the tile closed, and 3.0 (SD 2.8) instances where they selected a non-target item.

To validate that the manipulation of temporal cost of errors was effective, we compared the median times from error occurrence to recovery between the two error types, for each manipulated tile reopening time (Figure 4). Examining the box plots, we do not see a systematic difference in recovery time between the two error types.

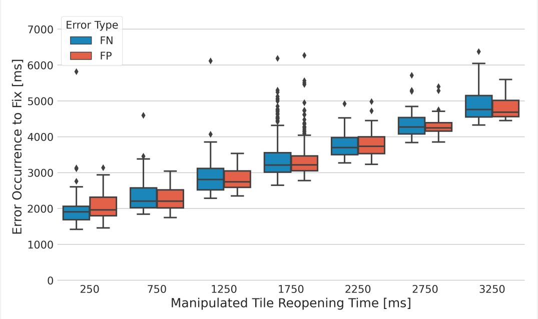


Figure 4. Box plot of median times from error occurrence to fix (one data point for each error type / time cost, per participant).

We also analyzed the difference in the medians of measured time to recover from FP vs. FN errors for each unique participant/cost-regimen (Figure 5). A regression line fit on this data is very close in intercept and slope to x = y, indicating that the tile reopening time was effective at manipulating the difference in recovery time. Specifically, recovering from FP errors took a median of only 50 milliseconds longer than FN errors (IQR 654ms).

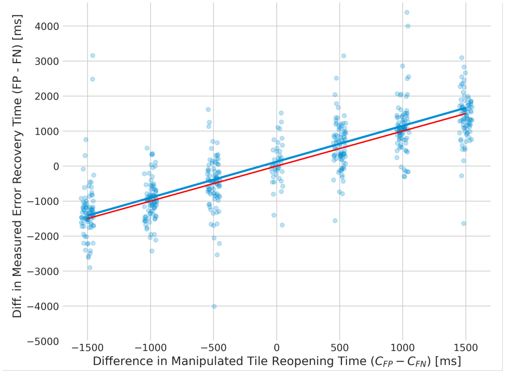


Figure 5. Difference in measured error recovery time between FP and FN errors (y-axis) vs. difference in manipulated tile reopening time between FP and FN errors (x-axis). Blue line: regression fit. Red line: x = y.

4.4.2 Error Preference

Following each condition, participants provided a forced-choice preference for an error type, followed by a strength rating for that preference. Coding responses with strength rating ("No preference") as neutral, participants preferred FN errors in 347/558 blocks (62.2%), FP errors in 187/558 blocks (33.5%), and were neutral in 24/558 blocks (4.3%). That FP errors were preferred in half as many blocks as FN errors even though reopening costs were balanced across conditions suggests a bias against FP errors that cannot be explained by the temporal cost of errors alone.

To quantify the bias against FP errors, a logistic regression model was fit to the response data, with preference for FP errors as the outcome variable, a fixed effect Reopen_Delta (the difference in tile reopening time between error types, see Section 4.3), and Participant_ID as a random effect (Figure 6 left). Reopen_Delta was found to be a significant predictor ($X^2(1)$=35.705, p < .0001). The *indifference point* – the value of Reopen_Delta at which the model predicts error preference will shift from FP to FN – occurs when Reopen_Delta is -1.55 seconds (95% confidence interval: [-2.9s, -0.6s]), suggesting a bias against FP errors equivalent to ~1.5 seconds of added recovery time.

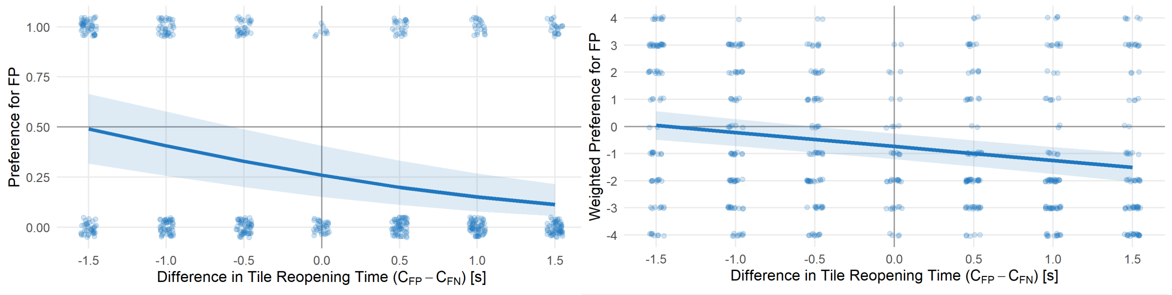


Figure 6. Error preference models. Left: mixed-effects logistic regression on FP preference; Right: Mixed-effects linear regression on strength-weighted FP preference. Data points jittered to show density. Error bands indicate 95% confidence intervals.

A weakness of the logistic regression model is that it does not consider the strength of preference ratings provided by participants. To address this, a weighted preference for FP errors was computed by multiplying the strength responses (0=No preference, to 4=Very strong preference) by error type preference (+1 for FP, -1 for FN), yielding values on a scale from -4 to +4. A mixed-effects linear regression model was fit with weighted preference for FP errors as the outcome variable, Reopen_Delta as a fixed effect, and Participant_ID as a random effect. Reopen_Delta was found to be a significant predictor (F(1,514.09)=41.948, p < .0001). Examining the model fit using this approach (Figure 6 right), the indifference point occurs at -1.43 seconds (95CI: [-2.6s, -0.5s]), indicating a bias against FP errors consistent with the logistic regression model above. For the remainder of error preference analyses in this paper, weighted preference for FP errors is used as the outcome variable, since it captures both the error preference and strength responses.

The above analyses indicate that the temporal cost of the two error types was a significant driver of error preference. In addition, the apparent bias against FP errors suggests that temporal cost is not the sole driver of error preference. In the coming sections we seek to understand what other factors may be influencing error preference, and to what degree.

4.4.3 Frustration

To investigate whether the temporal error costs associated with FP and FN errors are also driving frustration in the task, a linear mixed-effects model was fit with outcome variable *Frustration* (participants responses to the frustration question), fixed effects C_{FP} and C_{FN}, and Participant_ID as a random effect. The model indicated a significant positive effect of C_{FN} (F(1,513.00) = 9.96, p < .01), but not of C_{FP} (F(1,513.03) = 0.01, p = .92). Given the apparent bias against FP errors as a driver of error preference, this is surprising. It may be that the bias against FP errors is strong enough that the reopening cost associated with those errors was immaterial, at least for the set of costs tested in this experiment. In Study 2, a wider range of cost regimens is tested to investigate this further.

4.5 Post-Study Questionnaire Results

The post-study questionnaire asked participants: *In general, which type of recognizer error did you find to be more frustrating?* In response, 30/44 selected false positives, 11/44 selected "Both were equally frustrating", and only 3/44 selected false negatives. Participants were also asked to briefly explain what they disliked about each error type. An open-coding approach was used to identify common themes, which are reported in the sections below.

4.5.1 Common Themes for Both Error Types

Across both FP and FN errors, the most common rationale for disliking an error type was the added time to recover, or that the error impeded progress on the task (mentioned in 22/44 FP comments, 18/44 FN comments). Closely related, participants expressed that they disliked having to "backtrack" or duplicate effort (9/44 FP, 7/44 FN). Finally, participants expressed a dislike for the system acting counter to their intentions, ignoring a click they knew they had made, or acting as though they had clicked when they knew they had not (12/44 FP, 13/44 FN).

4.5.2 False Positive Specific Themes – Attention Cost, Effects of Task Context

For FP errors, five participants suggested that it was challenging to notice these errors when they occurred, e.g:

> *[FP] errors were more frustrating. I had look for the items getting selected for no reason. That was very annoying. – P162*
>
> *I felt like I was always on edge and anticipating one of these when I opened a box and it wasn't the green circle. I simply dreaded these more for whatever reason. I can't pinpoint why, but they are much more frustrating. – P143*

These comments suggest that FP errors force the user to be more attentive. This makes sense, because FP errors do not occur in response to the user intentionally performing an action. As a result, the user must expend cognitive resources to monitor for their occurrence. Related to this point, two participants mentioned an added effort to move the mouse back to the affected tile if they did not notice the error quickly, e.g.:

> *I disliked [FP errors] as I'd often have to move the cursor back to correct it because I didn't notice the error right away. – P152*

It is important to note that the design of the study task may have reduced the attention required for noticing FP errors, as compared to potential real-world systems, since FP errors only ever occurred immediately following the reveal of a non-target item. The effort required to identify the consequences of an FP error were also minimized, because the study system clearly indicated the tile where a correction must be made. Real-world systems are unlikely to have this type of self-awareness, potentially leading to more severe costs.

A second notable theme was that FP errors were seen as worse because they occurred on tiles that did not contain the target item, or conversely that FN errors were viewed less negatively because they occurred when the user has found the target item (mentioned by 10/44 participants, across both error types), e.g.:

> *[...] I think [FN errors] were far less annoying, partly just because psychologically, a [FN] still means you've located the green circle and are almost done with that part. It's like getting pushed back a foot from the finish line, whereas unintended selection [FP] is like getting pushed back to the starting line as soon as you take off. – P143*

It may be that the recovery actions for FNs are viewed less negatively because they are seen as contributing to progress on the overall task (by selecting a target item), whereas FPs are seen as additional work that could have been avoided. The following comments support this explanation:

> *Missed selection [FN] errors did not bother me that much since I knew I was going to have to select them anyway – P133*

> *I didn't like the [FP errors] because it just seemed like an extra step that was impeding my progress of only selecting circles – P137*

Another explanation is that the interruption of searching for target items to deselect a non-target item imposes a task-switching cost [14], which requires one to keep track of which tiles have already been checked while recovering from the FP error (imposing a load on memory and retrieval [13]). For FN errors, even though the recovery requires the same actions as for an FP, it may be viewed as part of the primary task as the user has no need to change their task goal.

A final potential explanation is that finding the target item acts as a reward, which partially offsets the frustration of the FN error that occurs immediately afterward.

4.5.3 *False Negative Specific Themes – Error Attribution*

Comparatively fewer themes stood out in participants' comments on FN errors. However, two participants made comments that suggest they partially attributed these errors to themselves:

> *It felt that it was my error that the box was not selected, only slightly less frustrating. – P154*

> *I didn't like that it slowed me down and not knowing if I clicked it correctly or not. – P168*

P154's comment makes it clear that they were aware that FN errors were not caused by them, even acknowledging that they were less frustrating for that reason, but still states that they "felt that it was my error". P168's comment suggests ambiguity about whether the user or the system has caused these errors. True to how FN errors typically manifest, participants detected FN errors primarily through an *absence* of confirmatory feedback (the user feels the feedback of clicking the mouse button, but the tile is not selected), which may have created ambiguity about whether the user or the system was responsible for the error. It is worth noting that the study system did provide some additional feedback to enable users to distinguish false negative errors – the cursor outline turns from black to grey for a short time. Though this feedback was subtle, it may be that were it not present, participants would more strongly attribute FN errors to themselves. However, further research would be needed to investigate this.

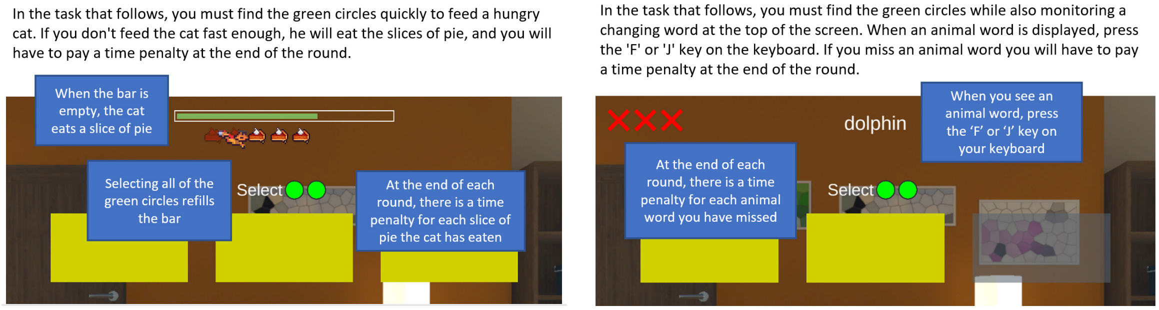
In summary, the questionnaire results are consistent with the quantitative analysis, demonstrating a bias against FP errors. Moreover, the comments suggest several potential explanations for this bias, and further insights into how these errors are experienced.

5 STUDY 2

The results of Study 1 suggest a bias against FP errors that may be explainable by added cognitive costs over FN errors. However, only one task was tested, and the tested range of temporal reopening costs did not extend out far enough to fully capture the switch in preference from FP to FNs. To address these limitations, and gain insights into how high-level task may influence relative error type preference, a second study was conducted.

5.1 Task Variations

Two variations on the *Standard* task from Study 1 were developed, which maintained the tile search task while adding additional elements (Figure 7).



In the task that follows, you must find the green circles quickly to feed a hungry cat. If you don't feed the cat fast enough, he will eat the slices of pie, and you will have to pay a time penalty at the end of the round.

When the bar is empty, the cat eats a slice of pie

Selecting all of the green circles refills the bar

At the end of each round, there is a time penalty for each slice of pie the cat has eaten

In the task that follows, you must find the green circles while also monitoring a changing word at the top of the screen. When an animal word is displayed, press the 'F' or 'J' key on the keyboard. If you miss an animal word you will have to pay a time penalty at the end of the round.

When you see an animal word, press the 'F' or 'J' key on your keyboard

At the end of each round, there is a time penalty for each animal word you have missed

Figure 7. Instructions for the study task variants. Time-Pressure variant (left); Split-Attention variant (right)

Time-Pressure (Figure 7a) – In addition to the standard task, a progress bar is displayed at the top of the screen, with an animated cat and five slices of pie. The progress bar counts down during the task, and if it becomes empty, the cat "eats" a slice of pie, the bar refills, and the process continues. After each page, there is a 4-second penalty for each slice of pie that has been eaten, the bar refills, and all slices of pie are restored. The idea is to add a sense time pressure to the task. The speed of the progress bar was tuned such that the penalties would be rare if participants worked quickly.

Split-Attention (Figure 7b) – In addition to the standard task, the user must monitor a changing word at the top of the screen. The user is instructed to press a key on their keyboard whenever they see an animal word. If they miss a word, an 'X' is added at the top left of the screen. After each page, there is a 4-second penalty for each 'X', and the 'X's are cleared. The idea behind this variant is to keep the user's attention split between the standard and secondary tasks, as might occur when the user's attention is not solely focused on providing input to a system.

5.2 Study Design and Procedure

The study followed a mixed design, with between-subjects factor TASK (*Standard*, *Time-Pressure*, *Split-Attention*), and within-subjects factor COST REGIMEN. Given that Study 1 indicated a bias against FP errors at the far limit of the differences between C_{FP} and C_{FN} that were tested, a new set of cost regimens (Table 3) was chosen to gather more data around the point where average preference switched from FP to FN in Study 1. Note that $C_{FN} > C_{FP}$ in 7 conditions, $C_{FN} < C_{FP}$ in 4 conditions, and $C_{FN} = C_{FP}$ in the remaining 2 conditions.

Table 3. Cost regimens tested in Study 2.

C_{FN} \ C_{FP}	0.25	1.00	1.75	2.50	3.25	4.00	4.75
0.25			✓				
1.00			✓				
1.75			✓				
2.50			✓				
3.25	✓	✓	✓	✓	✓	✓	✓
4.00			✓				
4.75			✓				

The study procedure was similar to that of Study 1. For the task variants, the set of practice blocks were conducted with the Standard task (i.e., without the extra elements of the task variant), followed by the task variant instruction page (Figure 7), and a final practice block with the task variant before the data gathering blocks began.

5.3 Participants

52 new participants were recruited for each of the three tasks, which took 60-75 minutes to complete. After filtering out participants who did not complete the study, and participants whose duration for the non-practice task conditions was greater than 3 standard deviations above the mean, or whose number of user errors was greater than 3*IQR above the median, 142 participants were included in the analysis (Table 4).

Table 4. Participant demographics for Study 2

	Standard Task (N=50)	Time-Pressure Task (N=49)	Split Attention Task (N=43)
Age	M = 34 (SD 10, range: 20 to 69)	M = 35 (SD 9, range: 23 to 64)	M = 35 (SD 10, range: 23 to 63)
Gender	38 Male, 12 Female	34 Male, 14 Female, 1 N/S	27 Male, 16 Female
Handedness	46 Right, 4 Left	42 Right, 7 Left	41 Right, 2 Left
Pointing Hand	49 Right, 1 Left	47 Right, 2 Left	43 Right, 0 Left
Computer Hw	20 Laptop, 30 Desktop	19 Laptop, 30 Desktop	23 Laptop, 20 Desktop
Pointing Device	50 Mouse	49 Mouse	43 Mouse
Gaming Frequency	Daily (23), Weekly (18), Monthly (4), Yearly or less (5), Never (0)	Daily (13), Weekly (18), Monthly (10), Yearly or less (5), Never (3)	Daily (18), Weekly (9), Monthly (6), Yearly or less (9), Never (1)

5.4 Results

Summaries of the three task conditions are shown in Table 5. In terms of average block duration, Standard (225s) is between the Time-Pressure (217s) and Split-Attention (241s). As in Study 1, the error injection was correct in most blocks (>97.7% for all tasks). In terms of task penalties (i.e., how often participants incurred a 4-second penalty for working too slowly in the Time-Pressure task, or missing words in the Split-Attention task), the Time-Pressure task penalties are in a tight range, suggesting the task was manageable. In contrast, the Split-Attention task has more penalties and a wide IQR, which may suggest that some participants had difficulty with the secondary task.

Table 5. Summary statistics for Study 2, by task condition

	Standard Task	Time-Pressure Task	Split Attention Task
Block Duration	M=225 (SD 121, min 139, max 2137)	M=217 (SD 83, min 136, max 986)	M=241 (SD 90, min 147, max 886)
Correct # of Injected Errors	637/649 (98.2%)	633/636 (99.5%)	543/556 (97.7%)
Task Penalties	N/A	Median=20 (IQR 13, min 8, max 88)	Median=33 (IQR 46, min 7, max 305)

Prior to the analyses that follow, the filtering criteria for participants and blocks from Study 1 (see Section 4.3) were applied, removing 8/142 participants who had confused the two error types, and 28/1841 blocks in which the correct number of errors was not injected.

5.4.1 *Error Preference*

Participants' weighted FP preferences by Reopen_Delta and task are shown in Figure 8. As in Study 1, the indifference point for each of the tasks occurs for Reopen_Delta values less than zero, suggesting a bias against FP errors. However, the magnitude of these estimates are lower than in Study 1, at around -0.4s for the Standard task (95CI: [-2.0s, 1.3s]), -0.2s for Time-Pressure (95CI: [-1.0s, 0.7s]), and -0.7s for Split-Attention (95CI: [-3.9s, 2.4s]).

This data was analyzed using a linear mixed-effects model with Reopen_Delta as a fixed effect and random effect Participant_ID. Additionally, to test differences between the task variants and the Standard condition, fixed effects were included for the two task variants (coded as binary indicator variables), and for interactions between Reopen_Delta and the task variant indicator variables. This enabled us to test whether the task variants show significant differences relative to the Standard condition, in terms of both the slope and the intercept of how weighted FP preference responds to Reopen_Delta. In terms of statistical tests, a significant effect of Reopen_Delta was found (F(1,1577.29)=38.51, p < .0001). A significant effect was also found for the slope (but not intercept) for Time-Pressure (F(1,1577.22)=12.98, p < 0.001), suggesting that error preference was more sensitive to differences in temporal error cost for this task, relative to Standard. No additional significant effects were found.

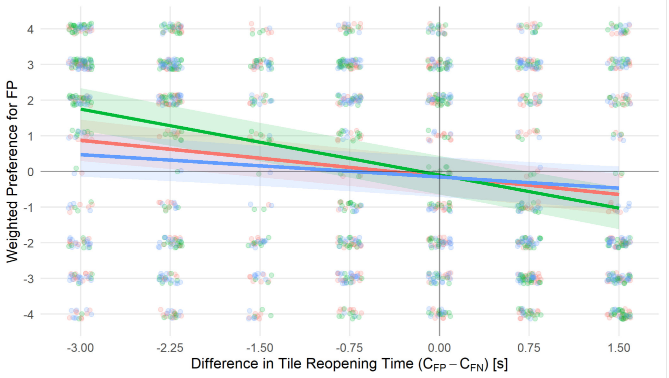


Figure 8. Strength-weighted error preferences for Study 2 (Standard, Time-Pressure, Split-Attention).
Error bands show 95% confidence intervals.

Given the wide range of task penalty counts across participants for the task variants, a follow-up analysis was conducted to see whether secondary task performance may be influencing error preference. Mixed-effect models with outcome variable weighted FP preference, fixed effects for Reopen_Delta and the number of task penalties, and random variable Participant_ID were fit for each task variant. For Split-Attention, the number of task penalties was found to be a significant predictor of weighted FP preference (F(1,40.39)=18.93, p < .0001). For Time-Pressure, no such effect was found. Based on this, Split-Attention participants were divided into two sub-groups – low-penalty participants (less than or equal to the median), and high-penalty participants (greater than the median).

Re-running the analysis of weighted FP preference by Reopen_Delta with Split-Attention divided, we see a more obvious difference between task groups (Figure 9). Participants in Split-Attention (low-penalty) exhibit a strong bias against FP errors, equivalent to ~2.7 seconds of added reopening time (indifference point at -2.7s, 95CI [-5.6s, -0.7s]). In contrast, participants in Split-Attention (high-penalty) exhibit little sensitivity to differences in reopening time between the error types. In terms of statistical test results (Table 6), we find a marginally significant effect for the Split-Attention (low-penalty) intercept (p = .064), and significant effects for the Split-Attention (high-penalty) slope and intercept. These results may suggest that the Split-Attention participants comprised two distinct groups – those who performed well on the secondary task, and those who were overwhelmed by the dual-task paradigm, leading to both poor performance on the word-identification task and difficulty with assessing error preference after blocks.

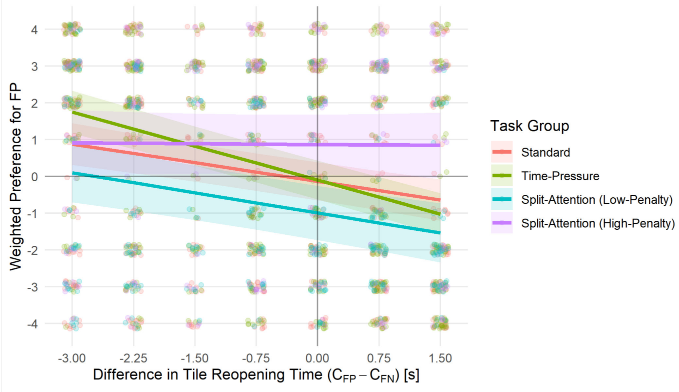


Figure 9. Strength-weighted error preferences for Study 2, with Split-Attention participants grouped based on task penalties. Error bands show 95% confidence intervals.

Table 6. ANOVA for mixed-effects model (outcome variable: weighted preference for FP errors)

Predictor		Estimate	F-statistic	p-value
Reopen_Delta		-0.337	F(1, 1576.31)=38.72	**<.0001**
Time-Pressure	slope	-0.280	F(1, 1576.23)=13.05	**<.001**
	intercept	+0.035	F(1, 135.34)=0.01	0.926
Split-Attention (low-penalty)	slope	-0.027	F(1, 1576.41)=0.08	0.778
	intercept	-0.858	F(1, 135.67)=3.50	0.064
Split-Attention (high-penalty)	slope	+0.322	F(1, 1577.17)=9.90	**<.01**
	intercept	+1.002	F(1, 136.08)=4.17	**<.05**

In summary, the analysis suggests that the Time-Pressure variant did not significantly affect error preference, but for participants who performed well on the Split-Attention variant, the bias against FP errors was higher. This is consistent with the qualitative comments from Study 1, which suggested that attentional demands are increased by FP errors, which could make this error type more frustrating in a task that demands more of the user's attention.

5.4.2 Frustration

To investigate the effects of temporal error costs on frustration, a mixed-effects model was fit with participants' frustration ratings as the outcome variable, C_{FP} and C_{FN} as fixed effects, and Participant_ID as a random effect. Both C_{FP}

and C_{FN} were found to be significant predictors (C_{FP} estimate +0.073, F(1,1578.2)=12.14, p <.001; C_{FN} estimate +0.055, F(1,1578.2)=6.99, p <.01). This suggests that the temporal cost of both FP and FN errors is a driver of frustration.

To analyze how frustration changed with error costs for the task variants, a linear mixed-effects model was fit with frustration ratings as the outcome variable, and a fixed effect *Reopen_Sum* (C_{FN} + C_{FP}) which captures the cost of both error types in one metric. Additionally, fixed effects were included for each task variant group (to capture their effect on the intercept compared to Standard), and for interactions between the task variant groups and Reopen_Sum (to capture their effect on the slope compared to Standard). Participant_ID was included as a random effect. A plot of the model results is shown in Figure 10, and statistical test results are shown in Table 7.

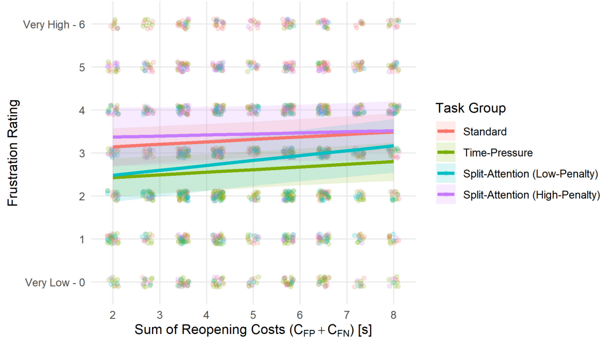


Figure 10. Frustration ratings by Reopen_Sum for Study 2, with Split-Attention participants grouped based on task penalties. Error bands show 95% confidence intervals.

Table 7. ANOVA for mixed-effects model (outcome variable: frustration ratings)

Predictor		Estimate	F-statistic	p-value
Reopen_Sum		0.058	F(1, 1576.27)=6.09	**<.05**
Time-Pressure	slope:	+0.004	F(1, 1576.18)=0.02	0.899
	intercept:	-0.723	F(1, 220.38)=4.39	**<.05**
Split-Attention (low-penalty)	slope:	+0.056	F(1, 1576.17)=1.86	0.172
	intercept:	-0.770	F(1, 220.66)=3.34	0.069
Split-Attention (high-penalty)	slope:	-0.033	F(1, 1576.53)=0.57	0.451
	intercept:	+0.294	F(1, 222.34)=0.43	0.515

Somewhat surprisingly, the model indicates lower frustration ratings for Time-Pressure and Split-Attention (low-penalty) as compared to Standard, with a significantly lower intercept for Time-Pressure (p < .05). There are a few possible explanations for why frustration might be lower for the Time-Pressure task. First, it may be that participants saw the errors and their costs as the mechanics of a game (trying to prevent the cat from eating the pie), rather than as aberrant events. Second, it may be that the time-pressure elements created a sense of reward when participants avoided penalties, which offset the frustration caused by errors. Finally, it may be that the time pressure pushed participants to focus on the task, and not engage in multi-tasking during the study (a known behavior for participants on Mechanical Turk) – the lower average and maximum block durations for the Time-Pressure task as compared to Standard may provide some evidence for less multi-tasking in the Time-Pressure task. The reward and multi-tasking explanations could apply to the Split-Attention (low-penalty) group as well, which also shows a trend toward lower frustration ratings.

None of the slope effects for the task variants were found to be significant, suggesting that the magnitude with which frustration increased with added cost was about the same across the task variants.

5.4.3 Exploratory Analysis of Additional Factors

An exploratory analysis was conducted to investigate additional factors that may influence preference for FP vs. FN errors. We were interested in whether error preference can be influenced by *primacy effects* (how close to the start of a block errors were experienced), *end effects* (how close to the end of a block errors were experienced), and *peak effects* (clusters of an error type in a short period of time). To test this, a set of metrics to capture these effects was developed (see Appendix). We also wanted to test the effects of user errors (where a user either opened a target item but failed to select it; or accidentally selected a non-target item). A mixed-effects model was fit with weighted error preference as the outcome variable, Reopen_Delta and the metrics above as fixed effects, and Participant_ID as a random effect (Table 8).

Table 8. ANOVA for mixed-effects model (outcome variable: weighted preference for FP errors)

Predictor	**Estimate**	**F-Statistic**	**p-value**
Reopen_Delta	-0.391	F(1, 1578.2)=127.10	**<.0001**
User Errors (failed to select target)	+0.013	F(1, 1687.3)=0.06	.810
User Errors (selected non-target)	-0.026	F(1, 1649.4)=0.14	.708
Primacy_Delta	+0.044	F(1, 1600.5)=0.57	.451
Peak_FP	-0.113	F(1, 1610.0)=4.21	**<.05**
Peak_FN	-0.110	F(1, 1623.4)=3.73	.054
End_Delta	-0.126	F(1, 1598.4)=4.54	**<.05**

The model results indicate a significant effect for Peak_FP (i.e., a peak of FP errors occurring during the task is associated with a lower preference for FP errors), and for End_Delta (i.e., the error type experienced more frequently near the end of the block is less preferred). These results suggest that retrospective assessments of the experience of recognizer errors can be influenced by recency, and that clusters of errors occurring together can stand out when making retrospective assessments. This has implications for study methodologies for understanding the user experience of errors, which are discussed at the end of this paper.

5.5 Post-Study Questionnaire Results

In the post-study questionnaire, participants were asked: *In general, which type of recognizer error did you find to be more frustrating?* For the Standard task, participants' responses were: FP=17/Equally-Frustrating=19/FN=14, for Time-Pressure: 19/11/19, and for Split-Attention: 14/18/11. In interpreting these responses, it is important to remember that the temporal cost of FP errors was greater than that of FN errors in only 4/13 (30.8%) of the conditions that each participant experienced, so the fact that FP errors are disliked as much as or more than FNs is further evidence of a bias against them not explainable by temporal costs alone.

In terms of participants' responses on what they disliked about each error type, the prevalent themes largely mirrored those for Study 1. However, two new themes emerged, as did some themes related to the Time-Pressure task variant.

5.5.1 Higher Average Cost for One Error Type

One new theme, mentioned by four participants in reference to FN errors, was that the tile reopening times seemed longer in general for the error type, e.g.:

> *I did not like the time spent waiting for the missed selection errors which felt to be on the longer side for much of the trials. … - P296*

This makes sense since there were in fact more conditions in which FN errors had higher reopening times. Surprisingly, two participants made similar comments for FP errors.

5.5.2 *Expectations from Prior Experiences*

Another new theme was to mention past experiences with a faulty mouse. Three of four participants who raised this theme mentioned that it made FN errors seem less severe, e.g.:

> *I actually had no problem with [FN] errors. I've had it happen from time to time in the past due to a malfunctioning mouse and I've gotten used to checking over my own work and clicking twice to ensure that my selection is counted. – P328*

One final participant mentioned this as a negative:

> *I think [FN errors] also has a negative connotation in my mind from an old mouse that I no longer use in which I would click and not get a response. – P296*

An additional participant, commenting on FP errors, mentioned that these errors never occur with mice:

> *I hated this one. Nothing more frustrating than having your mouse click when you don't actually click anything. I've never had my mouse do that before. – P265*

Though these comments were rare, they suggest a need to replicate the current study with other input modalities, to better understand how expectations from past experiences may be influencing participants' assessments.

5.5.3 *Task Variant Themes*

In terms of themes specific to the task variants, a theme that was raised for the Time-Pressure variant was that FN errors were frustrating because they could occur at critical moments – i.e., when the timer is about to run down. This theme was mentioned in one comment on FP errors, and four comments on FNs, e.g.:

> *[FN errors] often led to me losing a piece of my pie since, by the time I found the circle, I was already low on time. Therefore the frustration with this error was simply that it was more obvious how much it was hurting my time because I was already feeling the time crunch before I got there and was in a hurry to click the circle. – P267*

It is important to note that, objectively, both error types are equally damaging in the Time-Pressure task. However, when searching for the final target on a given page, FN errors can lead to a situation where the user *almost* avoids a penalty, but then receives one because of an error. These comments indicate that the specific circumstances of how errors occur in a task can influence the cost of that error.

Finally, a participant in the Time-Pressure condition commented that errors (of both types) were less frustrating because it was fun to try to beat the timer, which may partially explain the lower frustration ratings for Time-Pressure.

> *The reason why I was not frustrated or annoyed is because, it was sometimes fun, trying to beat the timer/bar, even when it clicked when you didn't, or didn't when you did. – P240*

In contrast to the Time-Pressure task, no obvious novel themes emerged in comments by Split-Attention participants.

6 DISCUSSION

In this section, the results of the two studies are compared, and their main findings are summarized and related to existing research. We also discuss the implications of these findings for the design of gesture-based user interfaces, and for further studies to understand the experience of system errors.

6.1 Differences in bias estimates for Study 1 vs. Study 2

While both studies showed evidence for a bias against FP errors, the bias estimate of ~1.5s in Study 1 is greater than the ~0.4s bias estimate for the Standard task in Study 2, which is surprising given that these two groups performed the same task. There are several potential explanations for this. First, the set of cost regimens tested in Study 2 meant that participants experienced more conditions where C_{FP} was less than C_{FN}, which may have led to a global decrease in negative sentiment against FP errors as compared to Study 1, where the differences in cost were balanced across the full set of cost regimens tested. Second, the set of cost regimens tested in Study 2 may have artificially reduced the apparent bias against FPs by gathering fewer "anchor" data points for conditions with $C_{FP} > C_{FN.}$ Future studies could address this by testing a wider range of cost differences. Finally, Study 2 used a coarser resolution in the set of costs that was tested (0.75s increments rather than 0.5s as in Study 1), which could have made the study less sensitive to detect a bias. Investigating these possibilities is an important area for future work to replicate the present results, and to develop robust and reusable methodologies for measuring biases in error-type preference.

6.2 What drives error-type preference?

Through controlling the recovery actions associated with FP and FN errors, we have demonstrated that temporal costs can drive both error type preference and the frustration associated with each of these error types. The study results also demonstrate a bias against FP errors not explained by temporal costs alone, and evidence for several potential causes of this bias.

A primary driver of the bias against FP errors appears to be the added cognitive cost of noticing when FP errors have occurred, as evidenced by the greater bias in the Split-Attention (low-penalty) group and post-study comments. Comments suggest that these errors were more surprising, and that they demanded the user's attention to notice their occurrence. Moreover, once noticed, the errors had to be corrected. Here, the user must switch their primary task goal from finding target items to undoing an erroneous selection of a non-target. This type of goal-set switching is known to impose cognitive costs, particularly if one does not expect the onset of the new task [14] and/or if it requires one to retrieve from long-term memory [13] (e.g., to remember the location of the most recent selection). A task-switching cost is consistent with Quinn's work on negativity biases in interactions, which showed that error recovery actions that feel like backtracking can lead to a more negative view of an assistive technique, recovery time being controlled for [17]. A final potential driver of the bias against FP errors is the lack of feedback associated with FN errors, which may lead to ambiguity about whether the error was the fault of the system or the user. The importance of feedback for error recovery is well-established [25], but more work is required to understand how ambiguity about the *attribution* of errors (by users to themselves vs. to the system) might affect relative error preferences and perception of input techniques.

In addition to the finding that the Split-Attention task may exacerbate the bias against FP errors, participants' lower frustration ratings and post-study comments on the Time-Pressure task suggest that task context can influence the experience of errors in both global ways (such as the Time-Pressure task exhibiting lower frustration, potentially because it prevented multi-tasking or because errors were seen as part of a game) and in more granular ways (such as where a user narrowly misses beating the timer because of an FN error).

The exploratory analysis of Study 2 data also provides evidence that retrospective assessments of error-type preference can be influenced by peak and end effects. This is consistent with Katsuragawa et al.'s finding that multiple recognizer errors occurring in succession are particularly damaging to user experience [10], and adds to existing research demonstrating that peak and end effects can influence retrospective assessments of interactive systems [4, 6].

6.3 Implications for design and further research

The findings in this paper have several implications for the design of recognition-based input systems. First, understanding the factors that drive error cost lays the groundwork for error-cost aware input recognizers, which dynamically assess the relative costs of FP versus FN errors, and adapt the recognizer threshold accordingly. A challenge in implementing such a dynamic thresholding approach is how to integrate in cognitive costs and biases (which are not observable) with observable costs (such as the temporal cost studied here, or other observable costs such as number of input actions). The studies reported here enable such an integration of cognitive costs by estimating those costs *in terms of* an observable cost. For example, the finding that users exhibit a bias against FP errors equivalent to ~1.5 seconds could be directly integrated into a dynamic approach to adjusting a recognizer threshold.

Second, to address the challenge that false positive errors do not occur in response to a user action, and thus require the user to monitor for their occurrence, an input system could help the user with noticing these errors. General techniques have been proposed to draw the user's attention to changes in an interface [2], as have techniques for notifying the user of changes in wide display spaces [3]. In the case of FP errors, there may be an opportunity to dynamically assist users with noticing errors, e.g. when scores are close to the recognizer threshold.

Finally, the studies reported here have implications for further research into the effects of errors on user experience. The finding that frustration caused by temporal cost of errors was lessened in the Time-Pressure task suggests that gamifying a study task to understand errors can result in an unrealistic picture of the effects of errors. The peak/end effects suggest that it may be valuable to control for these effects in studies of error cost. It also suggests value in developing means for measuring the effects of errors in real-time, to avoid relying on retrospective evaluations. Finally, the tile-opening task developed for this study provides a means for independently controlling the rate and temporal cost of FP and FN errors, and we hope it can be adapted for further research in this area.

6.4 Generalizability and limitations

An important open question is how the results reported here will generalize beyond mouse input and the specific study task that was tested. If we consider another input modality, such as mid-air gestures for AR/VR, we could imagine there being greater attentional demands to notice when FP errors have occurred, or to identify the effects of these errors, which may be outside the user's field of view. Given that our study results suggest that attentional demands are a driver of bias against FP errors, this may translate into a greater bias against FP errors, though more investigation is needed to test this hypothesis. Gesture-based input may also introduce additional costs that are not present for the mouse, such as the effort to prepare to perform a gesture (e.g., by moving the hand into a position where it can be sensed), and the effort to perform the gesture itself. Such additional costs could influence the ease with which users recover after FP and FN errors, and thus error-type preference, so it will be important to identify such additional costs and understand their effects.

Another important open question is how expectations and prior experience with an input modality influence error-type preference. Given that mouse input is highly reliable and familiar to users, it is possible that users' prior experiences may have influenced their experience of errors. If these effects are not symmetric across FP and FN errors (e.g., because FN errors with a mouse are more believable than FP errors), it could explain some of the bias against FP errors that we observed. However, the comparison of the Split-Attention (low-penalty) vs. Standard groups in Study 2 should have controlled for any such effects, and suggests that greater attentional demands increase bias against FP errors.

The question of generalizability to other tasks and error-rate regimens is important as well. The study task used in this work may have made the temporal cost of errors more salient because it was the only factor varying across blocks.

As well, the tile reopening visual draws attention to reopening time, and participants experienced many errors in a short time, enabling them to rapidly learn the temporal costs. Understanding how the temporal cost of errors influences error type preference in a broader range of tasks and situations is an interesting area for future work.

7 CONCLUSION

This paper has contributed new findings on how relative preference for false positive versus false negative errors is influenced by the temporal and cognitive costs associated with these errors. This is a first step toward more comprehensive models and understandings of the costs of input errors on user experience, and a vision of error-cost aware gesture recognition techniques that can dynamically adjust their behavior to prevent errors when they would be most costly, and create a more optimal user experience.

ACKNOWLEDGMENTS

The authors would like to thank Dan Clarke for developing the tile task component of the study system, and Thomas White for developing the remote logging infrastructure.

REFERENCES

[1] Nikola Banovic, Tovi Grossman, and George Fitzmaurice. 2013. The Effect of Time-based Cost of Error in Target-directed Pointing Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 1373–1382. https://doi.org/10.1145/2470654.2466181

[2] Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. 2006. Phosphor: Explaining Transitions in the User Interface Using Afterglow Effects. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (UIST '06), 169–178. https://doi.org/10.1145/1166253.1166280

[3] Anastasia Bezerianos, Pierre Dragicevic, and Ravin Balakrishnan. 2006. Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology* (UIST '06), 159–168. https://doi.org/10.1145/1166253.1166279

[4] Andy Cockburn, Philip Quinn, and Carl Gutwin. 2015. Examining the Peak-End Effects of Subjective Experience. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 357–366. https://doi.org/10.1145/2702123.2702139

[5] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. 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* (CHI '17), 278–288. https://doi.org/10.1145/3025453.3025739

[6] Carl Gutwin, Christianne Rooke, Andy Cockburn, Regan L. Mandryk, and Benjamin Lafreniere. 2016. Peak-End Effects on Player Experience in Casual Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 5608–5619. https://doi.org/10.1145/2858036.2858419

[7] Seongkook Heo, Jiseong Gu, and Geehyuk Lee. 2014. Expanding touch input vocabulary by using consecutive distant taps. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 2597–2606. https://doi.org/10.1145/2556288.2557234

[8] Ken Hinckley, Patrick Baudisch, Gonzalo Ramos, and Francois Guimbretiere. 2005. Design and analysis of delimiters for selection-action pen gesture phrases in scriboli. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '05), 451–460. https://doi.org/10.1145/1054972.1055035

[9] Eric Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '99), 159–166. https://doi.org/10.1145/302979.303030

[10] 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: 15:1-15:30. https://doi.org/10.1145/3181672

[11] 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 *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 347–356. https://doi.org/10.1145/2702123.2702603

[12] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 411:1-411:14. https://doi.org/10.1145/3290605.3300641

[13] Ulrich Mayr and Reinhold Kliegl. 2000. Task-set switching and long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26, 5: 1124–1140. https://doi.org/10.1037/0278-7393.26.5.1124

[14] Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3: 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

[15] George Nagy. 1982. 29 Optical character recognition—Theory and practice. In *Handbook of Statistics*. Elsevier, 621–649. https://doi.org/10.1016/S0169-7161(82)02032-X

[16] Matei Negulescu, Jaime Ruiz, and Edward Lank. 2012. A recognition safety net: bi-level threshold recognition for mobile motion gestures. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services* (MobileHCI '12), 147–150. https://doi.org/10.1145/2371574.2371598

[17] Philip Quinn. 2016. Economic behaviour and psychological biases in human–computer interaction. University of Canterbury, Christchurch, New Zealand. Retrieved April 16, 2019 from https://ir.canterbury.ac.nz/handle/10092/12187

[18] Gerhard Rigoll, Andreas Kosmala, and Stefan Eickeler. 1997. High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, 69–80.

[19] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '10), 2863–2872. https://doi.org/10.1145/1753846.1753873

[20] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1-8. https://doi.org/10.1145/3290605.3300750

[21] Jaime Ruiz and Yang Li. 2011. DoubleFlip: a motion gesture delimiter for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 2717–2720. https://doi.org/10.1145/1978942.1979341

[22] C. J. Van Rijsbergen. 1975. Evaluation. In *Information Retrieval*. Butterworth & Co., 95–132. Retrieved from http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html

[23] Bryan Wang and Tovi Grossman. 2020. BlyncSync: Enabling Multimodal Smartwatch Gestures with Synchronous Touch and Blink. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–14. https://doi.org/10.1145/3313831.3376132

[24] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM* 62, 6: 70–79. https://doi.org/10.1145/3282486

[25] Daniel Wigdor, Sarah Williams, Michael Cronin, Robert Levy, Katie White, Maxim Mazeev, and Hrvoje Benko. 2009. Ripples: utilizing per-contact visualizations to improve user interaction with touch displays. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology* (UIST '09), 3–12. https://doi.org/10.1145/1622176.1622180

[26] Jacob O. Wobbrock, Andrew D. Wilson, and Yang Li. 2007. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In *Proceedings of the 20th annual ACM symposium on User interface software and technology* (UIST '07), 159–168. https://doi.org/10.1145/1294211.1294238

[27] 2020. F1 score. *Wikipedia*. Retrieved January 22, 2020 from https://en.wikipedia.org/w/index.php?title=F1_score&oldid=935689499

A APPENDIX – FEATURES FOR EXPLORATORY ANALYSIS

Table 9. Features used in the exploratory data analysis reported in Section 5.4.3.

Feature	**Description**
User Errors (failed to select target)	Count of instances in the block where the user opened a tile to reveal a target item but did not click to select it before the tile closed.
User Errors (selected non-target)	Count of instances in the block where the user opened a tile to reveal a non-target item and then mistakenly selected it.
Primacy_FP	Sum of the timestamps of FP errors in the block (i.e., time since the start of the block), with an exponential weighting function $f(t)=10*e^{(-t/30000)}$ applied to each.
Primacy_FN	Sum of the timestamps of FN errors in the block (i.e., time since the start of the block), with an exponential weighting function $f(t)=10*e^{(-t/30000)}$ applied to each.
Primacy_Delta	Primacy_FP – Primacy_FN
Peak_FP	A gaussian kernel density estimate (bandwidth=5000) was fit to the timestamps of FP errors in the block. Metric is the sum of density estimates at each timestamp.
Peak_FN	A gaussian kernel density estimate (bandwidth=5000) was fit to the timestamps of FN errors in the block. Metric is the sum of density estimates at each timestamp.
End_FP	Sum of the deltas between timestamps of FP errors and the end of the block, with an exponential weighting function $f(t)=10*e^{(-t/30000)}$ applied to each.
End_FN	Sum of the deltas between timestamps of FN errors and the end of the block, with an exponential weighting function $f(t)=10*e^{(-t/30000)}$ applied to each.
End_Delta	End_FP – End_FN