More Information May Reduce Errors for Novice Users: Usability Testing and Redesign of the Square Register App

Lawton Pybus, Michelle Taub, Hallie Clark, and Caleb Furlough
North Carolina State University

The popular Square Register app was tested and redesigned for better initial performance using human factors methods. The research team interviewed both current and potential users who expressed concerns and questions about the nature of the app. We conducted further usability tests to evaluate weaknesses in the app interface and to guide redesigns. We created semi-functional prototypes to test the benefits of these design recommendations. Results were mixed, but partially support including additional information in the app for the benefit of novice users. Issues found during prototype use are discussed.

INTRODUCTION

Since the advent of mobile Internet-connected touchscreen devices, such as the iPhone in 2007, several software developers have released point-of-sale applications for smaller, more affordable platforms to enable transactions in a variety of environments. One of the foremost of these applications is the Square Register (hereafter referred to as "Square") by Square, Inc. Initially released in 2010, Square has become commonly used in restaurants, coffee shops, hair salons, and is often seen at events such as concerts, conventions, and festivals. In August 2012, Howard Schultz, CEO of Starbucks, announced that his company would be investing $25 million into Square, Inc., and would begin using the app heavily in its store locations. It is important, therefore, that a popular financial application be immediately understandable for novice users while retaining functionality to complete sales.

In the present study, we describe how we evaluated Square, developed and tested redesigns based on these analyses, and performed statistical analyses to compare the current app and our redesigns. We then describe our results and discuss the implication of our findings. We began with interviews of users and potential users, then conducted an environmental, heuristic, and task analysis of the app. Based on these usability tests, we created two alternative mock-up redesigns and tested them with participants.

METHODS

Environmental analysis. We performed an environmental analysis to evaluate how characteristics of the environment such as lighting, noise level, and ambient conditions like temperature may affect usability and errors.

Task analysis. To better understand each specific step the customer must complete in order to finish the transaction, we performed a task analysis using Square version 3.2.4 (released September 25, 2013) on an iPhone 5 running iOS version 7.2. These respective software versions were the most current versions available when the analysis was performed. We then analyzed each step for potential errors, and described the type of potential error (i.e. cognitive, perceptual, physical), the consequences of the error, and the criticality of the error. To evaluate the criticality of the error, we asked, "Would this error require a complete re-start of the transaction, or would it only delay its completion? If only a delay, how significant is the delay?"

Our task analysis included steps pertaining to the tip selection screen and signature screen, although vendors may elect to bypass these screens (for example, if tipping is not customary).

Heuristic evaluation. We conducted a heuristic evaluation to judge interface features against general usability standards developed by Jakob Nielsen and Rolf Molich (Dix, Finlay, Abowd & Beale, 2003).

Materials

User interviews. We administered a structured interview of open-ended questions with images of the app interface. Example questions included, "Are you familiar with the Square app?" and "What do you believe is the ultimate goal of this app?" We also asked about payment preference style (i.e. cash vs. credit).

Prototype mock-ups. Two alternative prototype mock-ups were developed using Pidoco and InVision. These alternatives are described in detail in the "Results" section, as their designs were based on the results of the above analyses and interviews. Pidoco is a web application for generating wireframe prototypes. In InVision, we assigned areas of the Pidoco wireframes to "hotspots," which upon tapping or clicking lead to another screen or action, creating a convincing semi-functional app prototype. Although the buttons were functional, other interface elements, such as keyboards and the signature screen, were not.

Post-test survey. We administered a 12-question questionnaire to participants after testing the prototypes, which included the System Usability Scale (SUS), the Network Promoter Score (NPS), and the Single Ease Question (SEQ). The SUS, developed in 1987 by John Brooke, was chosen because of its frequent use in usability studies as an indicator of perceptions of usability (Bangor & Kortum 2008). The NPS and SEQ are both single questions that address other perceptions of usability not directly measured by the SUS (Reichheld 2003, Sauro & Dumas 2009).

 DOI 10.1177/1541931214581235

Participants

User interviews. Nine volunteers were recruited from the greater Raleigh area. Four were male; five were female. Age was classified only as "younger" (younger than 25 years old, $n = 4$) and "older" (older than 60 years old, $n = 5$). Four participants were considered "experienced users," reporting having used the app at least once, and five were "novice users," having no prior experience with the app.

Prototype testing. We separately recruited 18 undergraduate and graduate student volunteers from North Carolina State University to test our prototypes, ranging in age from 19 to 37 years old ($M = 25.3$, $SD = 4.51$). 72% of participants were female. All participants reported having normal or corrected-to-normal vision and being moderately to extremely familiar with modern technology (such as smart phones and tablets, etc.). Eight participants (44.4%) reported no prior familiarity with Square, while three (16.7%) reported slight familiarity. Five (27.8%) were moderately familiar with the app prior to testing, and only two (11.1%) reported being "very familiar" with Square.

In both user interviews and prototype testings, participants took the role of "customer" rather than "vendor" as a user of Square.

Design

For prototype testing, participants were randomly assigned to test one of three prototypes: the original Square design and the two redesign mock-ups (one with an additional information screen and one without). The original Square interface was captured using screenshots and given "hotspots" with InVision, in a process similar to how the prototype mock-ups were made, so that equivalent comparisons could be made.

We used SUS, NPS, and SEQ scores to measure usability perceptions. Performance was measured by total time to complete the task and the number of errors (determined through task analysis) during performance. We expected that novice users (i.e. inexperienced or "slightly familiar" users) would demonstrate better performance and report better perceptions of usability when using our redesigns than the current Square design. Because one redesign included additional screens, we expected that overall task completion time would be longer, but that users would demonstrate fewer errors and better perceptions of usability than with the other prototypes.

Procedure

For prototype testing, participants were asked to give their informed consent and to then complete a demographic questionnaire about age, gender, vision, familiarity with technology, and familiarity with Square (see "Participants"). The conducting researcher gave some background information on the purpose of the study and provided an imaginary context for the task to be completed. Participants were to imagine that he or she were at a coffee shop and had just purchased coffee, tea, and a muffin using a credit card. The researcher, acting as the cashier, then handed the iPhone to the customer to complete the purchase. During testing, each participant's interaction with the iPhone was recorded on video so that performance could be timed and reviewed in detail later. After the task was completed to the participant's satisfaction, he or she completed the post-test survey described previously.

RESULTS

Interviews, Analyses, and Prototyping

User interviews. Novice and experienced users gave characteristically different responses about the nature of Square. Several novice users erroneously expected to interact with the software interface with a stylus or pen rather than by touch of the finger, notably on the signature screen. Similarly, several participants mistakenly believed the software would retain the credit card information beyond the transaction. One participant expressed hesitation to use the software without a detailed itemized receipt.

The results of these interviews focused our attention on differences between novice and experienced users (rather than between older and younger users), and suggested that additional information or instruction may help to ameliorate the performance differences between these groups.

Environmental analysis. Because Square is an app for mobile platforms (i.e. Apple's iOS, Google's Android), usage environments are practically limitless, so the analysis focused on locations where the application is commonly used (e.g. restaurants, coffee shops, hair salons, events). Each of these locations has important environmental characteristics that could affect usage. For example, a busy café with a long line of customers may contribute distracting noise and activity, and additional social pressure to perform quickly so that other customers may be served. Similarly, an outdoor event may create visibility issues due to glare on the screen.

Although these limitations would be difficult or impossible to overcome with interface design changes within our power to implement, they became the practical constraints of our recommendations. For example, because the environmental noise level may vary between silent and very loud, a proposed intervention would not include auditory instructions.

Task analysis. A recurring theme among the errors uncovered was the inability to correct an error without re-starting the transaction. For example, if the user mistakenly taps an undesired tip selection, there is no way to return to the tip selection screen to correct the error.

Another class of errors may be attributed to inadequate instructions or feedback. In the "custom tip selection" form on the tip selection screen, the software sets a maximum amount, ostensibly to prevent inadvertent over-tipping. This maximum amount varies depending on the sale amount, and is visible prior to entering an amount, but not afterwards. If the user inputs an amount in excess of the maximum, it automatically changes to the maximum without explanation. In other words, if the user intended to leave a tip exceeding the maximum amount without noticing the maximum limit beforehand, there is no feedback in the software to explain the change.

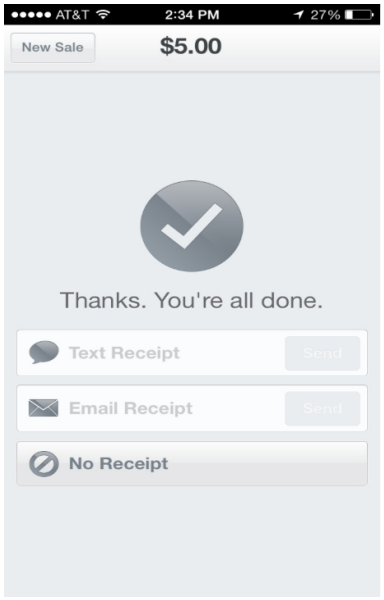

Figure 1. Receipt selection screen.

Finally, another group of errors may be attributed to improper, confusing, or misleading instructions and feedback. One typical example of this occurs on the receipt selection screen (see Figure 1). Although further action is required of the user, the interface indicates the contrary with a checkmark symbol and the feedback, "That's it! You're all done." Indeed, during testing, one participant using the original Square app prematurely ended the task on this screen, presumably due to the misleading feedback.

For each group of errors described, the examples given are typical of multiple errors uncovered during the task analysis, which are not listed in full for lack of space.

Heuristic evaluation. Much of the evaluation reinforced the importance of correcting issues indicated by previous analyses, but also revealed additional usability issues. For example, Square violates the convention of "consistency and standards." Buttons in the top left and right of the interface vary from screen to screen, and even within the same screen in one instance. Figure 2 shows the signature screen before and after the user has signed. In this screen, the top left button changes from "Cancel" to "Clear"—an unexplained and potentially unexpected change.

The app also violates the convention of "error prevention," which advises to "[e]ither eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action" (Dix, Finlay, Abowd & Beale, 2003) Users can neither correct mistakes by returning to the screen on which they occurred, nor can they review their input and choices in a confirmation dialog.

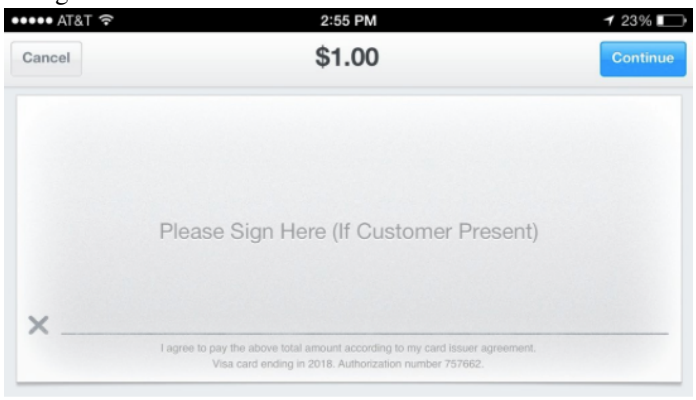

Figure 2. Signature screen, before and after user signature.

Prototype mock-ups. Two alternative redesign prototypes were developed, as described previously. The first included a "welcome screen" to inform novice users about the nature of the app and an "itemized receipt screen" with details about the transaction. These screens address concerns mentioned in the user interviews conducted, and were intended to prevent errors identified in the task analysis through instruction. We also adapted navigation features within the extant Square interface for consistency. "Back" and "Continue" buttons were placed in the same place across all screens, and functions like "Cancel" and "Clear" were placed in distinct buttons that did not change as they did in the original interface (see Figure 3). Feedback was also changed. On the tip selection screen, the maximum tip amount was always displayed next to the custom tip selection form. On the receipt selection screen, the checkmark symbol was changed to a receipt icon, and the text was changed to say, "Wait! Please choose a receipt option."

Because we expected the two additional screens to increase overall task completion time, we created a separate prototype that omitted these screens but maintained the other design changes applied in the first prototype.

Prototype Testing

Tip selection and receipt selection choices were left to participant discretion. To determine whether these choices may have had an influence on results, a cross tabulation of the occurrences of all choices was performed: χ^2 (8) = 14.2, p = .077 for all tip selections and χ^2 (4) = 7.3, p = .121 for all receipt selections. This suggests that receipt and tip choices were sufficiently varied.

Means and standard deviations for each measure by prototype and familiarity are provided in Table 1.

Errors. Some errors were attributed to limitations of the InVision software, such as glitches, hotspots that were too small, and non-functional keyboards. Other errors were attributed to the participant, likely due in part to misunderstanding the app or what actions were necessary to further the task. Another type of error was attributed to oversight in the construction of the prototype (i.e. a missing button). However, because this error occurred only once (representing .006% of all errors), it was coded as an InVision error.

Table 1. Descriptive statistics for errors, SUS, NPS, SEQ scores, and task time by prototype and familiarity*

	Redesign without additional screens		Redesign with additional screens		Original Square app	
	Novice users	Experienced users	Novice users	Experienced users†	Novice users	Experienced users
InVision errors	19 (26.87)	7 (26.87)	4.4 (3.36)	4	5 (1.63)	2 (2.82)
Other errors	6 (8.49)	3.5 (4.51)	3.6 (3.44)	3	.75 (0.96)	1.5 (0.89)
SUS score	63.75 (22.98)	76.25 (5.95)	62.5 (10.3)	80	90 (15.4)	95 (3.53)
NPS score	1 (1.41)	6.5 (2.38)	5.6 (3.78)	9	8 (2.16)	9 (1.41)
SEQ score	3.5 (4.95)	6.5 (0.58)	6.2 (1.79)	9	6.5 (1)	7 (1)
Total task completion time (seconds)	86.5 (75.66)	50 (12.52)	78.8 (26.42)	36	31.75 (3.2)	24 (1.41)

*All values are given in the form: *M* (*SD*). †Only one participant occurred in this condition.

Errors attributed to InVision represented 68.6% of all errors, while errors attributed to participants represented the other 31.4%. The two kinds of errors were highly related, $r(16) = .66$, $p = .003$, suggesting that participants experiencing a large number of errors due to the prototype software were also likely to make other errors. To test our hypotheses, we use only the errors not attributed to InVision in the following results.

Novice users made the most errors with the redesign without additional screens ($M = 6$, $SD = 8.49$), and the least with the original Square ($M = .75$, $SD = .89$). Novices using the redesign with additional screens made an average of 3.6 ($SD = 3.44$) errors over the task.

Perception of usability. Novice users' responses on the SUS score favored the original Square app most ($M = 90$, $SD = 15.4$), followed by the redesign without additional screens ($M = 63.75$, $SD = 22.98$), and the redesign with additional screens ($M = 62.5$, $SD = 10.3$). For both the NPS and SEQ, novice users preferred the original Square app the most, but preferred the redesign with additional screens to the redesign without them (see Table 1 for means and standard deviations).

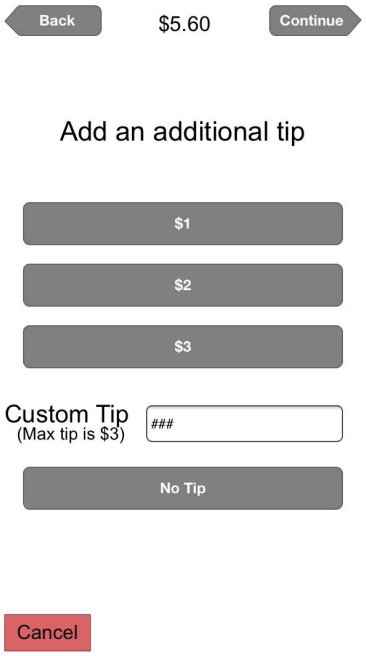

Figure 3. Redesign tip selection.

Time. Novice users took an average of 86.5 seconds ($SD = 75.66$) to complete the task in the redesign without additional screens, an average of 78.8 seconds ($SD = 26.42$) to complete the task in the redesign with additional screens, and an average of 31.75 ($SD = 3.2$) seconds to complete the task in the original Square app.

Discussion and Limitations

Due in part to the small sample size used in prototype testing and unequal conditions based on the participants' prior familiarity, these results were mostly reported without inferential statistics.

The measures used for errors, perception of usability, and time generally favored the original Square design. We suspect that this unexpected result could be explained, at least in part, by the prototypes used. Our alternative prototypes were developed in Pidoco, while the InVision reproduction of the Square app was created from screenshots. Thus, interface elements were different in appearance and size from Square elements, which were the same as those used in their functioning commercial product. It is possible, therefore, that comparisons should not be made from the original interface with the Pidoco-developed interface. The sizes of buttons and other interface elements may have influenced task completion times. Aesthetic preference for the original interface may have influenced perceptions of usability. Unfortunately, perceptions of aesthetics were not measured during the study.

InVision was chosen because of its ease of use and availability, but the major source of errors in this study was the prototype software itself. Furthermore, the relationship between InVision errors and other errors was troubling. Further testing should include higher fidelity prototypes, or should control better for these kinds of errors.

Despite these limitations, novice participants using our redesign with additional information screens ultimately demonstrated fewer errors than those using the redesign without them. Furthermore, they generally preferred the redesign with additional information to the one without. Interestingly, they even completed the purchasing task in less time with the additional screens, which we initially expected would increase the overall task time. One possible explanation is that the additional time taken to read the informational screen helped users to better understand the software, reducing time-consuming errors on later screens, and overall frustration.

Square, Inc. was uninvolved with the present study. However, during the course of this study, Square, Inc. coincidentally released a major update to the Square app—which included several interface changes addressing issues mentioned here.

ACKNOWLEDGMENTS

Special thanks to Dr. Anne McLaughlin for her comments and guidance throughout this research.

REFERENCES

Bangor, A., Kortum, P.T., & Miller, J.T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction,* 24, 574-594.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (2003). *Techniques for evaluation. Human-computer interaction (2nd edition).*

Reichheld, F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12), 46-54.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-taks usability questionnaires. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1599-1608. ACM.