Assignment M5

Michael Tong
mtong31@gatech.edu

Abstract. Assignment M5 performs and analyzes the result of a survey for the implementation of a new feature to improve the efficacy of folders, as well as the implementation of new buttons to increase the efficacy of responding to meetings in Outlook. A summary discusses future improvements to both evaluations.

Introduction

The overall purpose of this effort is to improve the interface of the Microsoft Outlook application, which has been narrowed down in Assignment M2 (Tong 2018a) as a goal to increase the efficacy of users from a work productivity standpoint. Assignment M4 (Tong 2018b) establishes three prototype evaluations and this report discusses the result of two of their executions. A qualitative evaluation is performed in the form of a survey to analyze an improvement to the folder categorization, and an empirical evaluation is performed for the potential implementation of a new button through an experiment.

Qualitative Evaluation

From the Assignment M2 survey results, it had been stated that improvements to the folder system are warranted, which led to the prototype and now evaluation of a new interface function. Here, we explore the ability to send emails from a folder which will then place correspondence back to that folder as opposed to the inbox, allowing users to autonomously organize their work. Execution of this evaluation was done in the

form of a survey which attempts to identify the acceptance of this option from a hypothetical perspective. In total, 39 responses were captured through the Georgia Tech peer survey platform where ten questions were asked to the student body. The survey and population did not change throughout the process but for future execution, the study hopes to include members outside of the student population, especially at an office where the use case was conceptualized for.

In comparison with the original survey, the number of participants has decreased drastically from 150 to 39, but the results follow certain similarities. Crucially, the new survey responses still show an emphasis on using Outlook for a working environment, which was the basis for this prototype.

To summarize the raw categorical results, 87% (34) of respondents reported that they use Outlook for work, 33% (13) for academia, and 10% (4) for personal. 77% (30) use the folder feature to organize emails, 61% (24) organize specific ones, while 21% (8) organize all of them. 79% (31) are content with the folder system but 85% (33) believe that the proposed feature will be useful. Plots of this data are shown in Figures 1 through 5 in the appendix.

What immediately stood out from the survey was the deliberate query of questioning if users were content with the folder system, then seeing if they would find the new feature useful. Implicitly, this means that users have not thought too much about how the system can be improved. Unfortunately, the high number of content responses may lead to the feature being unused since it requires a change in behavior while already satisfied, which is unlikely. On the bright side, this is coupled with a high positive response rate for the implementation of the redesign.

The analysis then led to the number of users using the folder system, which was assumed to be high given the emphasis on work productivity and the few mentions of it in the original M2 survey. What is found is slightly shocking, where nearly 80% of respondents state they use the folder system. While the evaluation did not specify a particular threshold expectation, the personal expectation was closer to 50%. From

the responses it's clear that the folder system is an integral part of the interface, which increases the motivation for this redesign.

While a large portion of respondents acknowledge they use the folder system, there is an equally large portion that are content with the interface. This draws some concerns such as the one mentioned in the previous paragraph, but also lends itself to the question of if the redesign is necessary and/or worthwhile. After analyzing a few of the free text responses for what specific improvements respondents may like to see in the interface, most desires are for autonomous folder organization such as AI integration. While this redesign does not fully establish a traditional learning paradigm, it may potentially act as a precursor or evaluation rule for an AI algorithm in the future. Additionally, it does share autonomous characteristics in the email interface.

Unfortunately, while the redesign description did interest a majority of individuals, the following free response section detailing dissatisfaction revealed a few flaws. Most of the negative responses highlighted that this feature requires the user to check the folder for the new email as opposed to being able to see it in the compiled inbox list. This is an important remark because it had not been considered during the evaluation period. To alleviate this concern the redesign is being adjusted slightly. Instead of immediately moving the email to the folder from which it was sent, the email will be copied into the folder and remain in the inbox, but instead of staying indefinitely, the message will be removed from the inbox after a short period of time after it has been read.

Empirical Evaluation

The empirical evaluation discusses another implementation to the interface which may improve the efficacy of Outlook in a working environment. Here, the study looks to improve the efficacy of responding to meeting requests. As detailed in Assignment

M4, the goal of this evaluation is to identify whether or not the implementation of additional buttons to respond to meetings and events as accept, tentative, or decline, with a dedicated button that appears on emails when highlighted. The experimental group used the existing delete, archive, and flag responses to simulate accept, tentative, and decline respectively. Users were sent meeting emails containing the specific action to perform. The control group utilized whatever method of responding to a meeting they felt comfortable with. Execution of this experiment was performed with eight participants, each doing the experiment ten times across the two approaches (control and experimental). To prevent the sequence of performing either the control or experimental first having an effect on the experimental results, the participants were split in half where each group performed either feature first. Administering the experiment led to a few hiccups and the process was adjusted to compensate for the issues.

Since the experimental procedures had significant changes, they will be discussed first. The first prominent issue was the tolerance of the timestamp. The metadata that is attached to each email when it arrives in an Outlook mailbox only tracks the message to the nearest second, and the procedures that this analysis attempts to differentiate require less than ten seconds to accomplish. This is problematic as the granularity may not be sufficient to accurately separate distinguish between the two scenarios. Next, it became apparent that the metadata also does not track when a response is performed, such as deleting an email, making it impossible to determine when a response was made within the system through this approach. This issue is circumvented by using a manual stop watch and overseeing the process, which adds a small error for reaction time, which may be negligible due to the first issue of granularity.

Administering the experiment utilized eight individuals with varying levels of proficiency. This discrepancy is not factored into this particular experiment due to the limited number of participants and ambiguity of the measure, but it should be noted as it is an important influence. Additionally, the experiment contained a small number

of participants which increases the effect of this possible disparity. Raw experimental results are found in Table 1 of the appendices.

To summarize the results, the t statistic and p value is calculated for each approach, as well as the mean, median, and standard deviation. Supporting the alternate hypothesis, it is found that the t-statistical value is large at 7.878, with a significantly small p value of 1.694E-11. Unfortunately, these hypothesis testing methods will not be completely accurate due to the right skewed nature of both distributions. From these values, there is significant evidence to believe that the results are genuine and not the result of happenstance, allowing us to reject the null hypothesis. Investigating deeper, the experiment finds that the new feature on average reduces the response time of a meeting by over a second from 2.77 seconds to 1.75 seconds. Additionally, the experiment also finds that the new feature is a bit more consistent in the response time rate, shown by the decrease in the standard deviation from 0.612 to 0.536.

Following a summary of the overall results, each person is also analyzed to determine if there was an improvement in speed individually. Figure 6 in the appendices shows that every participant has improved their response times when using the new interface, which is phenomenal news. Unfortunately, the conclusions must consider the limited sampling size per person, which was five per approach, totaling only ten data points each. As a result, the standard deviation shown in Figure 7 displays significant differences between people and methods.

While the experiment and analysis were relatively successful, numerous changes could be implemented for improvement. During the first few moments of the experiment where the metadata was not formatted as expected could have been addressed earlier if a more detailed evaluation is performed. In lieu of this approach, a simple stopwatch can be used since human reaction time should not cause a significant effect in terms of differentiating the two experiments. Additionally, with regard to the participant proficiency, this could have been better monitored either by having users sell assess their familiarity of the interface, or sample individuals who state that they use Outlook on a frequent basis. A larger study group would have also

been helpful in reducing the resulting variation. Another significant improvement that may be performed in future evaluations is the randomization of which approach to use when responding to the meeting (e.g. randomize the new and old method's when they are sent as opposed to sending them as groups of five). This was a slight oversight and was recognized too far into the experiment. The last improvement considered is the creation of a more formal arrangement to perform the experiment. The use of the delete, archive, and flag buttons are not an ideal alternative to the anticipated buttons which induces additional error. If the icons can be altered to mirror the expected buttons, as well as tie the responses to metadata with millisecond tracking, the experiment would be more successful.

Evaluation Summary

From the resulting evaluations, the next iteration of the design life cycle is unique to each consideration. Follow this analysis, it appears that the qualitative evaluation for the introduction of an additional option to send emails from a folder and have responses return to the same folder, should return to the design alternatives segment of the cycle. For the empirical evaluation of incorporating additional buttons to improve the efficacy of meeting responses, a reevaluation should be performed incorporating the improvements discussed in the section.

Beginning with the qualitative evaluation, additional needfinding needs to be performed to assess how useful the implementation of this feature would be. While the survey results are highly positive for the feature, the fact that most respondents are content with the system and would like to have a much more automated system in replacement of the folder system, raises doubts about the actual widespread use. Additional needfinding can be performed to assess what alternatives may exists which improve the folder system to the respondent's desires.

The desires of users can be compiled into a new needfinding experiment to assess the feasibility of an automated system for email organization. A system such as this may incorporate the proposed feature but would not be the sole feature released in the redesign. As such, an all-new design may be required to meet demand.

With these considerations, a future evaluation for an autonomous system would utilize a different approach due to the high degree of fidelity required for implementation. Since a major factor for assessing shared autonomy is how invisible the system is, an ideal evaluation would be empirical, and measuring how accurate the system is at categorizing emails into their respective folders via the experimenter.

For the empirical evaluation, the major questions that arises is how accurate and important the information gathered is for the redesign incorporating the new buttons. Further needfinding can improve the accuracy of the results with a more robust experimental set up, as well as increasing the number of participants and randomizing the sequence of questions.

While this evaluation did not produce any design alternatives, a higher fidelity evaluation can be implemented by improving experimental set up to include functional buttons for the meeting responses, as opposed to simulating them with the current interface.

However, before considering this increased fidelity prototype, an evaluation of how important is this improvement? Are users interested in saving a few seconds and clicks with the redesign? A survey or interview should be implemented to analyze the response of users to these results. The next evaluation should therefore assess the importance of these findings to identify if they are worth implementing. A few seconds and clicks may not appear to be significant, but some frequent users may appreciate the reduced cognitive requirements.

References

1. Tong, M. (2018a). Assignment M2. *OMS CS6750 Human- Computer Interaction*. Washington, DC.
2. Tong, M. (2018b). Assignment M4. *OMS CS6750 Human- Computer Interaction*. Washington, DC.

Appendices

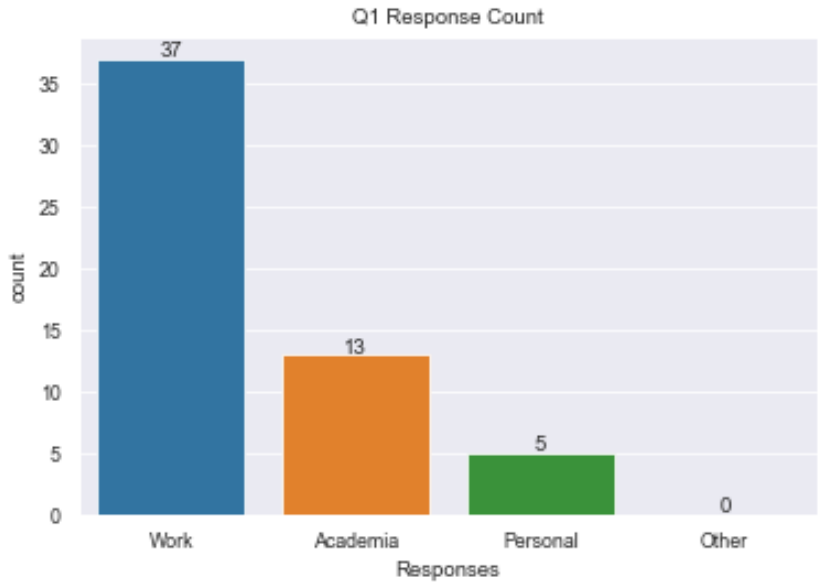

Figure 1: Question One Response Count

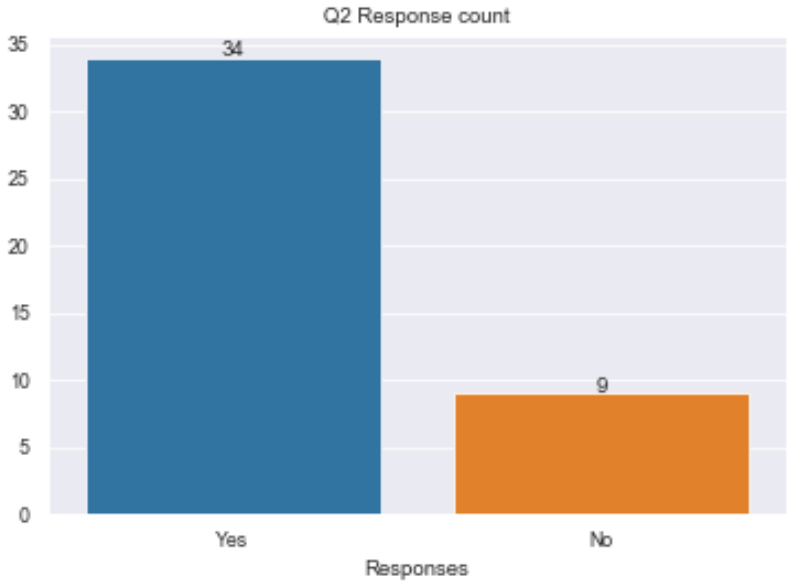

Figure 2: Question Two Response Count

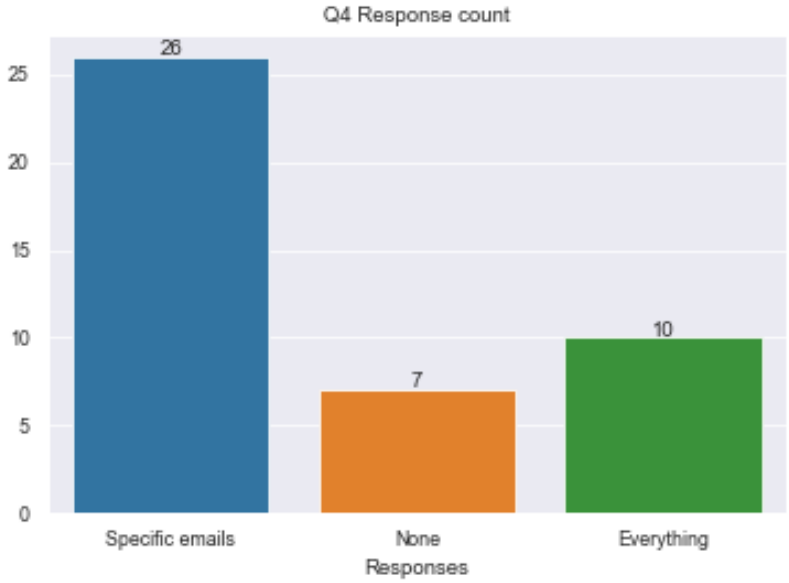

Figure 3: Question Four Response Count

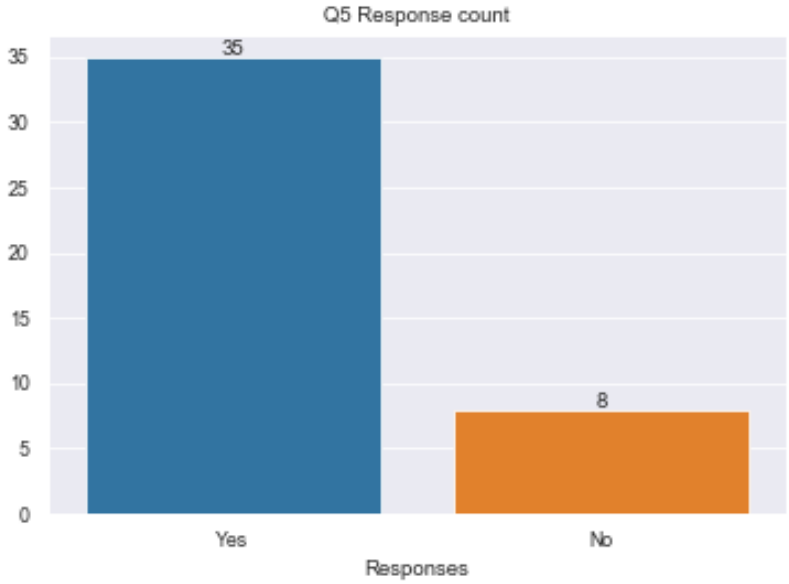

Figure 4: Question Five Response Count

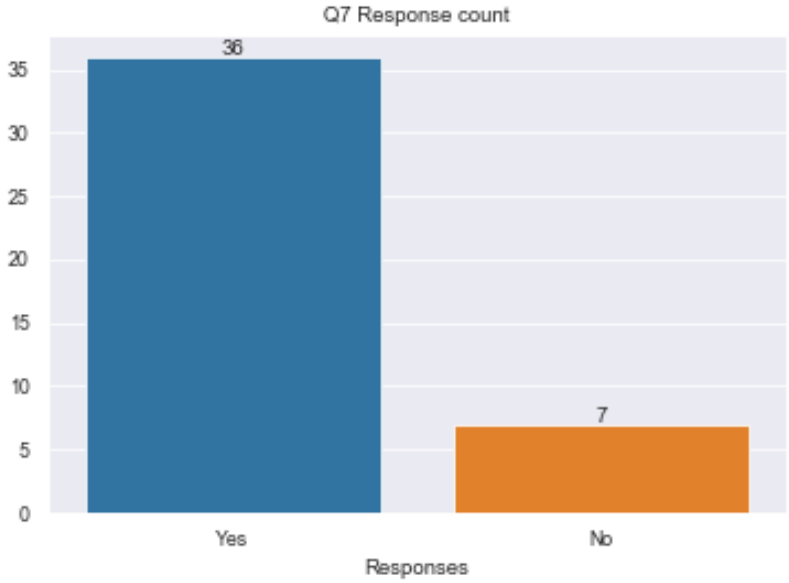

Figure 5: Question Seven Response Count

Person	Sent	Response	Method	Delta
1	19:58:54	19:58:56	N	0:00:02
1	19:59:19	19:59:20	N	0:00:01
1	20:01:33	20:01:35	N	0:00:02
1	20:04:46	20:04:47	N	0:00:01
1	20:06:30	20:06:32	N	0:00:02
1	20:10:20	20:10:24	O	0:00:04
1	20:11:33	20:11:37	O	0:00:04
1	20:13:11	20:13:14	O	0:00:03
1	20:14:05	20:14:08	O	0:00:03
1	20:15:54	20:15:57	O	0:00:03
2	21:06:16	21:06:18	N	0:00:02
2	21:08:07	21:08:09	N	0:00:02
2	21:09:44	21:09:46	N	0:00:02
2	21:10:55	21:10:56	N	0:00:01
2	21:13:25	21:13:27	N	0:00:02
2	21:16:01	21:16:03	O	0:00:02
2	21:17:28	21:17:30	O	0:00:02
2	21:19:38	21:19:41	O	0:00:03
2	21:20:58	21:21:00	O	0:00:02
2	21:21:33	21:21:35	O	0:00:02
3	18:07:49	18:07:51	N	0:00:02
3	18:08:41	18:08:43	N	0:00:02
3	18:10:32	18:10:34	N	0:00:02
3	18:11:50	18:11:52	N	0:00:02
3	18:14:57	18:14:59	N	0:00:02
3	18:16:49	18:16:51	O	0:00:02
3	18:19:36	18:19:38	O	0:00:02
3	18:22:50	18:22:53	O	0:00:03
3	18:24:35	18:24:37	O	0:00:02
3	18:25:51	18:25:53	O	0:00:02
4	18:35:49	18:35:52	N	0:00:03
4	18:36:41	18:36:43	N	0:00:02

4	18:38:32	18:38:34	N	0:00:02
4	18:39:50	18:39:52	N	0:00:02
4	18:42:57	18:42:59	N	0:00:02
4	18:44:49	18:44:51	O	0:00:02
4	18:47:36	18:47:39	O	0:00:03
4	18:50:50	18:50:53	O	0:00:03
4	18:52:35	18:52:38	O	0:00:03
4	18:53:51	18:53:54	O	0:00:03
5	19:18:49	19:18:52	O	0:00:03
5	19:19:41	19:19:45	O	0:00:04
5	19:21:32	19:21:35	O	0:00:03
5	19:22:50	19:22:53	O	0:00:03
5	19:25:57	19:26:00	O	0:00:03
5	19:27:49	19:27:51	N	0:00:02
5	19:30:36	19:30:38	N	0:00:02
5	19:33:50	19:33:52	N	0:00:02
5	19:35:35	19:35:36	N	0:00:01
5	19:36:51	19:36:52	N	0:00:01
6	20:24:49	20:24:53	O	0:00:04
6	20:25:41	20:25:44	O	0:00:03
6	20:27:32	20:27:35	O	0:00:03
6	20:28:50	20:28:53	O	0:00:03
6	20:31:57	20:32:00	O	0:00:03
6	20:33:49	20:33:52	N	0:00:03
6	20:36:36	20:36:38	N	0:00:02
6	20:39:50	20:39:51	N	0:00:01
6	20:41:35	20:41:36	N	0:00:01
6	20:42:51	20:42:52	N	0:00:01
7	20:45:59	20:46:02	O	0:00:03
7	20:46:31	20:46:33	O	0:00:02
7	20:47:40	20:47:43	O	0:00:03
7	20:49:36	20:49:38	O	0:00:02
7	20:50:36	20:50:39	O	0:00:03
7	20:51:58	20:52:00	N	0:00:02

7	20:53:40	20:53:41	N	0:00:01
7	20:55:55	20:55:57	N	0:00:02
7	20:57:56	20:57:57	N	0:00:01
7	20:59:40	20:59:41	N	0:00:01
8	19:11:56	19:11:59	O	0:00:03
8	19:12:41	19:12:43	O	0:00:02
8	19:14:54	19:14:56	O	0:00:02
8	19:16:49	19:16:52	O	0:00:03
8	19:17:31	19:17:34	O	0:00:03
8	19:18:50	19:18:52	N	0:00:02
8	19:20:36	19:20:37	N	0:00:01
8	19:22:33	19:22:35	N	0:00:02
8	19:24:30	19:24:32	N	0:00:02
8	19:25:55	19:25:57	N	0:00:02

Table 1: Empirical Data Raw Data

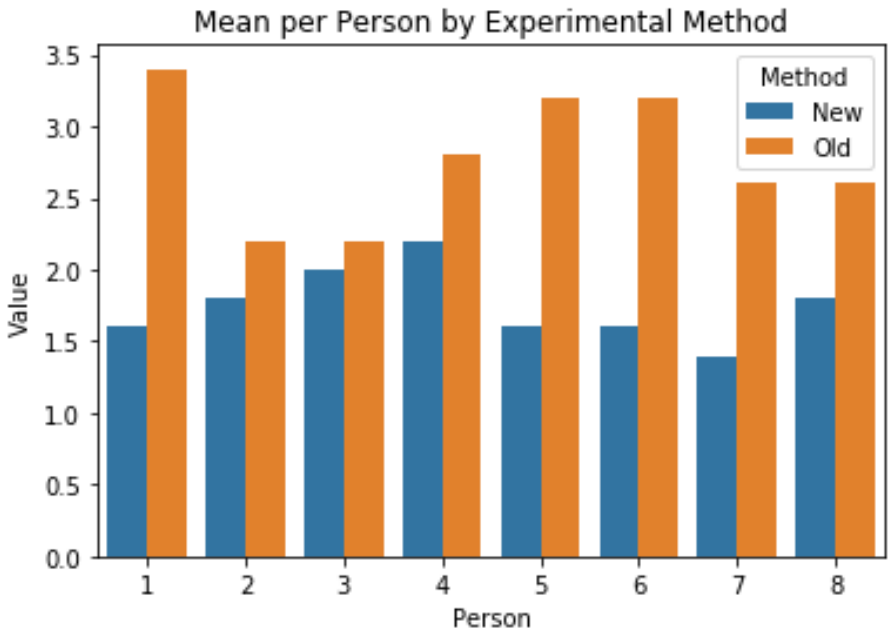

Figure 6: Empirical Evaluation Mean per Person

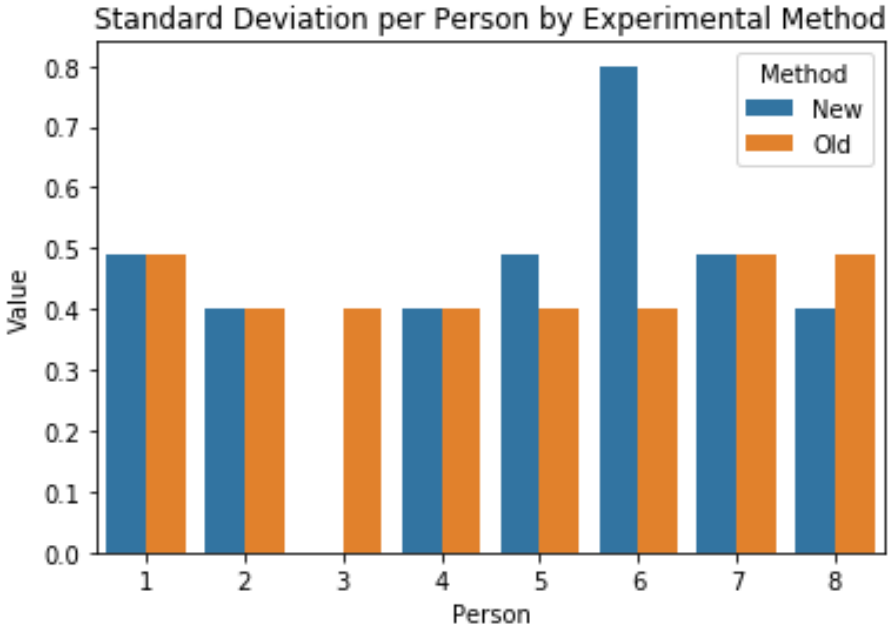

Figure 7: Empirical Evaluation Stdev per Person