Does the type of the interface agent alerting potentially unsafe emails affect the mean duration of fixation on the agent of participants?

Wider study context

Information technology has become a vital part of contemporary society. Many financial services and social interactions are directly dependent on the use of internet and numerous digital services it provides. With such a large number of users and personal information available online, internet thus has created an ideal environment for fraudsters to prey upon a large number of users. Phishing email is one of the methods used by fraudster to deceive users into providing them with sensitive personal or professional information. The fraudsters utilise phishing emails to deceive their victims by pretending to be a message from a reputable source and thus tricks their victims into providing the fraudsters with sensitive personal information, such as login id, passwords, or sensitive banking information etc. There are various security systems available for the users to help them assist in avoiding such phishing email attacks, however even sophisticated security systems will never be fool proof as the user always remain as a weak link. This is further shown in the study by Pfeffel *et al.* (2019) that, their best performing phishing email was successfully able to fool up to 40 percent of the participants in their study, thus suggesting that the users would always be a vulnerability in preventing phishing attacks.

As the user is the vulnerability the phishers are trying to exploit, a method to reduce the risk of email phishing is to aid the user in identifying phishing emails. Organisations often hold regular training sessions for their students and employees to identify potentially unsafe emails and messages as a preventive action to avoid the phishing problem. A more interactive solution would be providing the user with a warning when opening an email, to remind them of potentially unsafe behaviours. In this study, a static interface agent of different emotional expressions is implemented into a simulated mailbox, acting as warning, and alerting potentially unsafe user behaviours. This work thus aims to explore the user's security consciousness with the implementation of an interface agent in the mailbox. An interface agent is a character with human-like face, designed to assist user in a task. An example of an interface agent is Office Assistant more commonly known as "Clippy", an animated character based on a paperclip with human-like facial features, designed to assist and help users working in Microsoft Word.

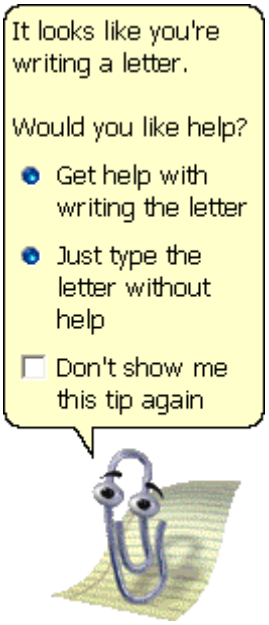

Fig1. Screenshot of "Clippy" in Microsoft Words 2000 (Wikipedia, 2021).

Tobii Pro Lab (2022) define fixations as periods of time where the velocity of a sequence of raw gaze points is under the defined velocity threshold, thus suggesting that the eye movements have remained relatively stationary. Potentially, such a long fixation on an interface agent could indicate that the participant is paying more attention to the agent while performing a task. Therefore, the participant's eye movements can provide a deeper insight on how the participants views and interacts with the simulated mailbox and the interface agent. Thus, to explore the effects of the interface agent, the research utilises Tobii TX300 eye-tracker to capture eye movements of the participants while they interact with interface agent.

Data provided

In this study, there are three experiment conditions, each condition has an interface agent implemented on the right side of the mailbox, with the right side being the area of interest. In the first condition, which is the control group, an image of a mailbox as an interface agent is given. In the second condition, a static interface agent with a neutral expression is used. And lastly, in the third condition, a static interface agent with a warning face is utilised. Before the start of the experiment, the participants perform a series of eye-tracking calibrations with the screen, this also helps in familiarising the participants with the simulated test mailbox. Once the calibrations are complete, the eye movements recording are then started, and the access to the simulated mailbox is provided to the participants. The participants are given a maximum of 30 minutes to remove any suspicious emails from the simulated mailbox. Once the participants deem, they have removed all the suspicious emails, the experiment is completed, and the participants fill in a post-experiment questionnaire. It is made sure that each participant can only participate in just one of the experiment conditions.

This study has recorded a set of two Excel files which contains data of eye movements captured with Tobii TX300 eye-tracker, and answers to post-experiment questionnaire. The post-experiment questionnaire data contains a recording number, participant number, demographics of the participants, the experiment conditions, how the participants perceived the agent rated based on Likert scale, and which expression the participants thinks is shown on the agent's face. The eye movements file contains data on the recording number, the experiment condition, time to first

fixation, duration of the fixation, mean duration of the fixation, count of the fixation, visit duration, visit count of both the on area of the agent and on the rest of the screen in seconds.

This dataset allows the investigation on the participants' mean duration of fixation on the agent in different experiment conditions, by comparing the means of mean fixation duration of the three unrelated groups. The mean of fixation duration on area of interest is the measurement of the average time of fixation on the interface agent, by dividing the sum of fixation duration of area of interest by the fixation count in the area of interest. The mean fixation duration on the area of interest allows comparison of the variable between the experiment conditions, without the influence of varying fixation count and time spent on the task of each individual participants.

Hypotheses

From the data provided, the independent variable of "experiment_condition" and the dependent variable of "fixation_duration_StimulusAOI_Mean", can be used to investigate if the type of the interface agent alerting potentially unsafe emails affects the mean duration of fixation on the agent of participants.

And thus, the hypotheses can be formulated as follows:

H0: The type of interface agent does not affect participant's mean duration of fixation on the agent.

HA: The type of interface agent affects participant's mean duration of fixation on the agent.

Analytical approach

To investigate the effects of the three experiment conditions on mean duration of fixation on the agent, the means of experiment conditions are compared. Before comparing the means of the experiment conditions, the assumptions for the test for the comparison of the means must be satisfied. For that, the data must satisfy the following assumptions: dependant variable should be of interval in nature, independent variable should be categorical independent groups, the observations should be independent, no significant outliers should be present, depending on the selected test the dependent variables should be normally distributed for each category of independent variable, and lastly, the homogeneity of variances should be present (Laerd statistics, 2018a).

As the independent variable "experiment_condition" is of categorical in nature and the dependent variable "fixation_duration_StimulusAOI_Mean" is of interval in nature thus the first assumption of the test is satisfied. As the participants of each experiment condition are different, the independent variable "experiment_condition" is considered as an unrelated and independent observation. To check for the outliers in the data, a boxplot can be utilised to graph the mean and spread of the data, with outliers being shown as a data point outside the spread of the boxplot. A test for normality and homogeneity of variances are also utilised to verify if the rest of the assumptions for comparison of means are satisfied.

Planned contrasts is a comparison between specific groups that were of interest before the collection of data (Laerd statistics, 2018b), the groups that are compared can be a combination of more than one independent group. As the study is exploring the differences between the three experiment conditions, with one of the conditions being the control group, a planned contrast can be performed to compare the means between the specified groups. For planned contrast, only two groups can be compared at each contrast, with the singled-out group in each contrast to be removed from subsequent contrasts. In each planned contrast the weights must equate to zero, while one

comparison group must be assigned to negative weights, the other compared group as positive weights, groups that are not involved would be given no weights (Field, 2018).

- **Rule 1**: Choose sensible contrasts. Remember that you want to compare only two chunks of variation and that if a group is singled out in one contrast, that group should be excluded from any subsequent contrasts.
- **Rule 2**: Groups coded with positive weights will be compared against groups coded with negative weights. So, assign one chunk of variation positive weights and the opposite chunk negative weights.
- **Rule 3**: If you add up the weights for a given contrast the result should be zero.
- **Rule 4**: If a group is not involved in a contrast, automatically assign it a weight of zero, which will eliminate it from the contrast.
- **Rule 5**: For a given contrast, the weights assigned to the group(s) in one chunk of variation should be equal to the number of groups in the opposite chunk of variation.

Fig2. The rules of planned contrast using weights (Field, 2018).

For this hypothesis, two contrast are performed. The first contrast compares control group of mailbox agent to the experiment groups with the agent with their means combined, to examine the differences of having an interface agent on mean fixation duration. The second contrast only compares the experiment group with neutral face and experiment group with warning face, to examine the differences of neutral agent and warning face agent on mean fixation duration.

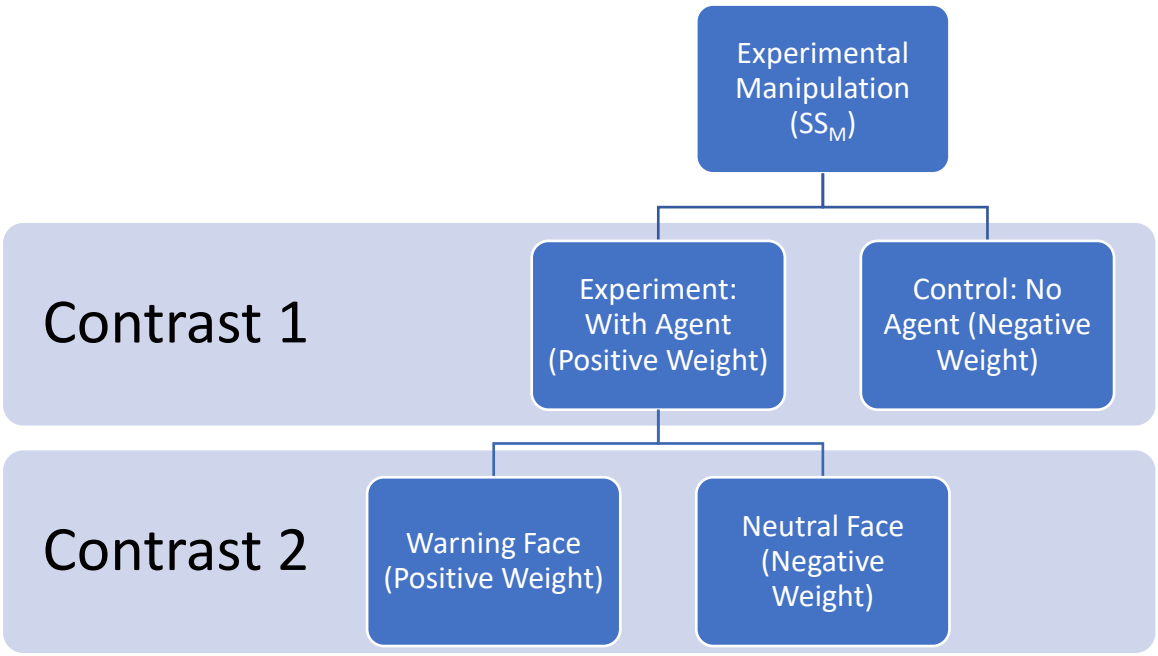

Fig3. The planned contrast

Tests selected

To compare the means of the three unrelated experiment conditions, if the data is normally distributed a one-way analysis of variance (ANOVA) is an appropriate method of comparing means. However, if the data is not normally distributed or if one of the variables is of ordinal type, the Kruskal

Wallis test is the appropriate test for the comparison. The independent variable “experiment_condition” is nominal, and the dependent “fixation_duration_StimulusAOI_Mean” is an interval data, which meets the assumption of comparison of means. Since the sample size is below fifty samples, Shapiro–Wilk test would be appropriate test for normality. Levene's test is used to verify the homogeneity of variances of the independent groups.

Once the assumptions are verified, depending on the normality of the data, ANOVA or Kruskal Wallis test can be performed. A planned contrast would be performed to evaluate the differences between experiment groups. In the first contrast, control group would be given -2 as weight, while each agent group would be given the weight of 1. In the second contrast, the neutral face group would have the weight of -1, the warning face group would have the weight of 1. If the results are significant and differences of variances are not significant, a Tukey's HSD post-hoc test can provide a deeper pairwise comparison between groups and assessing their significance of differences.

Test results

The data collected is first visually inspected and the means and spreads of the different experiment conditions are graphed using a box plot. The figure 4 below shows the plotted box plot of the collected data for different experimental conditions.

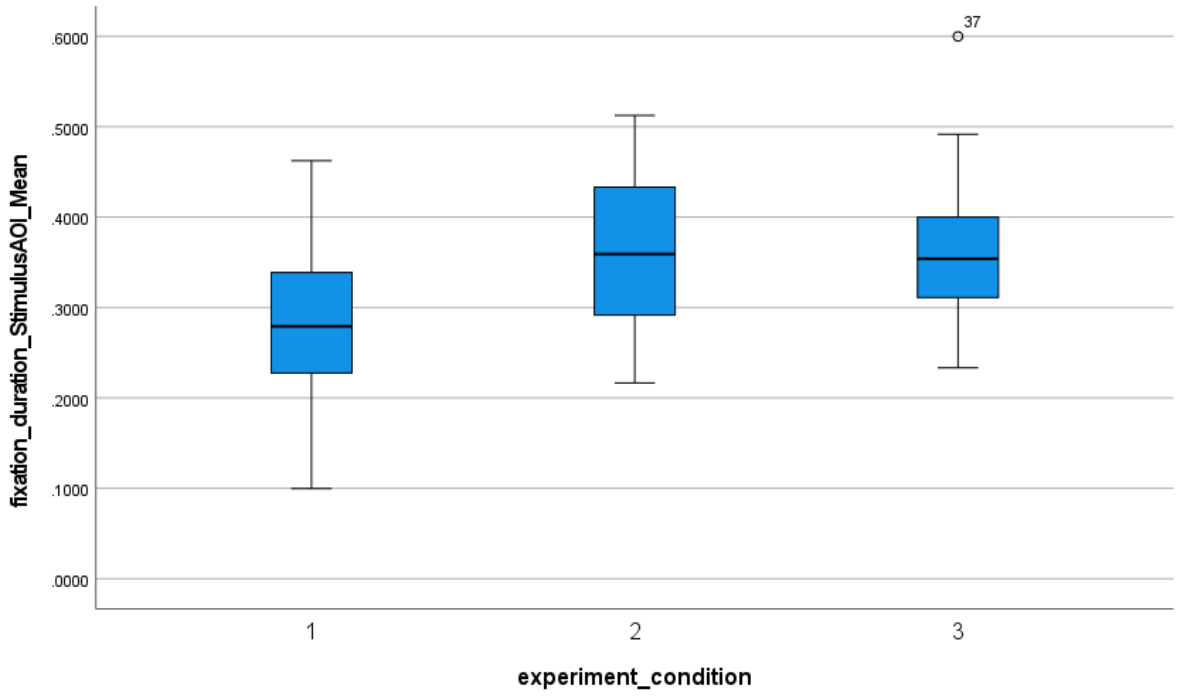

Fig4. Boxplot of collected data for each experiment condition.

Right away it can be observed that the experiment condition 1 which represents the control condition, has the largest spread in all the tested conditions and furthermore has visibly lower mean compared to the other conditions and the small size of the box shows that 50 percent of the collected data is quite close to the mean value. However, the large spread here shows that the collected data has quite a larger range compared to other categories. On the other hand, both the categories 2 and 3 namely, neutral face and warning face respectively, has very similar mean and range. Although smaller box in warning face category shows that 50% of values are very similar to the mean. It is also

noticeable that Record 37 is an outlier in the warning face category, with a mean duration of fixation of 0.60. This value is excluded in the statistical analysis conducted henceforth.

To verify the assumptions necessary for comparing means of different categories, a test for normality of the collected data is conducted. As the number of samples in the collected data are less than 50, the Shaprio-Wilk test for normality is used to determine the normality of the data.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	experiment_condition	Statistic	df	Sig.	Statistic	df	Sig.
fixation_duration_StimulusAOI_Mean	1	.123	12	.200*	.973	12	.941
	2	.132	11	.200*	.960	11	.778
	3	.177	12	.200*	.970	12	.916

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Fig5. Results from the Tests of Normality.

As shown in figure 5, all experiment conditions in this study are normality distributed as their significance value (p-value) are all greater than 0.05 (p= 0.941, p=0.778, p= 0.916 respectively).

With the condition of normality of the collected samples verified, the Levene's test for homogeneity of variances is conducted.

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
fixation_duration_StimulusAOI_Mean	Based on Mean	.450	2	32	.642
	Based on Median	.455	2	32	.638
	Based on Median and with adjusted df	.455	2	31.243	.638
	Based on trimmed mean	.432	2	32	.653

Fig6. Results from Tests of Homogeneity of Variances.

As shown in the figure 6, all the variances between each experiment conditions are not significantly different, this is evident as the significance value for all groups are greater than 0.05 (p based on mean = 0.642, p based on median = 0.638, p based on median adjusted = 0.638, p based on trimmed mean= 0.653). Thus, with the normality of data and the similarity of variances verified, the tests for comparison of means of different experiment conditions is conducted.

ANOVA

fixation_duration_StimulusAOI_Mean

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.049	2	.025	3.155	.056
Within Groups	.250	32	.008		
Total	.299	34			

Fig7. ANOVA results.

As all the data in the three tested categories are normally distributed and the variances of the data are not statistically significantly different to each other, and the number of groups being tested are greater than 2, the one-way analysis of variances (ANOVA) test is selected as the appropriate test for the comparison of the means. As evident in the figure 7, the significance value (p) in the ANOVA test is greater than 0.05 (p= 0.056), therefore the differences between the groups are significant. Although, the ANOVA test shows that there exists a difference in mean between the different tested groups, however how these groups differ from one another is explored utilising a priori orthogonal planned contrast test. For the first contrast, the experiment condition 1 i.e., the control conditions are selected as the first group and the conditions 2 and 3 are selected as the other group. For the second contrast conditions 2 and 3 are compared against each other.

Contrast Coefficients

	experiment_condition		
Contrast	1	2	3
1	-2	1	1
2	0	-1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	95% Confidence Interval Lower	95% Confidence Interval Upper
fixation_duration_StimulusAOI_Mean	Assumes equal variances	1	.156820	.0629610	2.491	32	.018	.028573	.285068
		2	-.014404	.0368891	-.390	32	.699	-.089544	.060737
	Does not assume equal variances	1	.156820	.0651413	2.407	20.485	.026	.021144	.292497
		2	-.014404	.0359929	-.400	18.857	.694	-.089776	.060969

Fig8. Planned Contrast results.

As shown in figure 8, the contrast 1 between control and agents group is significantly different (p=0.018) to each other as the significance value is less than the alpha value of 0.05. However, the Contrast 2 shows the differences between neutral agent and warning agent to be not significant (p=0.699) as the significance value is higher than the alpha value of 0.05. The difference in the means between the groups is further explored in depth utilising a post hoc Tukey HSD multiple comparison test.

Multiple Comparisons

Dependent Variable: fixation_duration_StimulusAOI_Mean

Tukey HSD

(I) experiment_condition	(J) experiment_condition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.0856121	.0368891	.067	-.176262	.005038
	3	-.0712083	.0360783	.135	-.159866	.017449
2	1	.0856121	.0368891	.067	-.005038	.176262
	3	.0144038	.0368891	.920	-.076247	.105054
3	1	.0712083	.0360783	.135	-.017449	.159866
	2	-.0144038	.0368891	.920	-.105054	.076247

Fig9. Tukey HSD results.

The figure 9 showcases the results from the Tukey HSD post hoc test. As the significance value for all the comparisons between the groups are greater than 0.05, this thus means that the differences between the groups are not statistically significant. Interestingly, while the planned contrast showed that while there exists a difference between the control group and the agent groups combined, however, according to Tukey HSD there does not exist a significant difference between the individual groups. Although, the difference between the control condition and neutral face agent while not statistically significant it is only so marginally as p= 0.067.

Interpretation

The Tukey test shows that the differences between the tested groups are not significant, however, planned contrast shows that the contrast between control and agent group is significantly different. This suggests that an interface agent influenced the mean fixation duration of participants, but the effect of neutral or warning face agent are not significantly different from each other, and not significantly different when comparing each agent group against control. Although the significance value between control experiment condition and neutral face experiment condition is 0.067 in Tukey's HSD, suggesting that the significance of the differences is only marginally not present.

Removing the single outlier is removing 2.56 percent of the samples, thus a substantial portion of data is excluded in the tests. This is evident when comparing the results of ANOVA (with outlier) and ANOVA (without outlier). The significance value of ANOVA has changes from not significantly different (p= 0.047) when keeping the outlier to significantly different (p= 0.056) when excluding the outlier. However, Tukey test suggested all groups are not significantly different with or without the outlier. And planned contrast continues to show contrast one being significantly different both with and without the outlier. Based on the conducted tests, the results suggest that there exists a need for greater number of samples. However, based on present data, it can be concluded that there is no statistically significant difference between all the tested groups.

Conclusions

On comparing the means of "fixation_duration_StimulusAOI_Mean" between the control, neutral face agent and the warning face agent groups, it was found using a one-way analysis of variance (ANOVA), that there exists a significant difference between the means of the groups. However, on conducting a post hoc Tukey HSD test, it was found that the differences between the groups are not

statistically significant. Based on this, it can be concluded that there is no statistically significant difference between all the tested groups.

Additionally, the study utilises fixation as a metric to determine the effectiveness of attracting user's attention to the agent. However, fixation is unable to determine the cognitive performances of the users. Perhaps, using a brain electroencephalogram data along with Tobii eye tracking data would provide a deeper understanding on the effects of the agents on the users. As longer fixation duration can also imply that the users are spending more time comprehending the information.

References

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. 5th ed. Los Angeles: Sage Publications.

Laerd statistics (2018a). *One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics*. [online] Laerd.com. Available at: https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php.

Laerd statistics (2018b). *One-way ANOVA in SPSS Statistics - Understanding and reporting the output.* [online] Laerd.com. Available at: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics-2.php.

Pfeffel, K., Ulsamer, P. and Müller, N.H. (2019). Where the User Does Look When Reading Phishing Mails – An Eye-Tracking Study. *Learning and Collaboration Technologies. Designing Learning Experiences*, 11590, pp.277–287. doi:10.1007/978-3-030-21814-0_21.

Wikipedia (2021). *Office Assistant*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Office_Assistant.