Investigating online information disclosure before sign-in Skype

Abstract

This paper is investigating the individuals of providing correct information during sign-in Skype app. A total of 900 respondents for asking several personal information, concern of privacy, knowledge and experience about Skype.

Respondents are randomly assigned into two groups, namely mandatory and voluntary. Then half of the mandatory and voluntary sample were allocated to the group that providing explanation the reason for Skype in collecting information. The remaining were assigned to the group that do not provide explanation. The order of delivery of survey was randomized for each respondent.

Respondents were also asked for their subjective perceptions of privacy concern, personal experience and knowledge on Skype. Putting these questions allows us to study how objective information proposed by the investigator reports to the subjective perception by the respondents, and to evaluate the tendency for their willingness to disclosure personal information in the surveys.

In summary, Mandatory group with explanation are more likely to provide correct information. In contrast, voluntary group without explanation are probably to give incorrect information. We also found that sensitivity, relevance and trust of questions requested have played significant effects of information disclosure. Furthermore, there is a weak correlation toward Skype knowledge and user experience. In order to get more correct information from survey, organization should first build trust relationship with users.

Introduction

Microsoft attained Skype in 2011. Skype can be used for multiparty communications inside calling with high-definition (HD) video, instant messaging and participation in meetings from various platforms.

According to the US FTC [1]stated that around 99% of online corporation will collect user personal information when they step in their websites. Supremely, corporation can use customer information for its targeted advertising practice and promotion.

In this study, customer information has been obtained by survey before sign-in procedure. We desired to identify and recognize the customer behaviours whether they are willing to provide correct information during two treatment measurements. Respondents are randomly assigned into two units, namely mandatory and voluntary. Then half of the mandatory and voluntary sample were allocated to the group that providing explanation the reason for Skype in collecting information. The remaining were assigned to the group that do not provide explanation. The order of delivery of survey was randomized for each respondent. The purpose of this paper is to explore the influence under two different treatments that affecting an individual who intent to disclosure their personal information to the organization.

Sample Data

The data in this study was using an online survey acquired before sign-in Skype. The findings can identify the tendency for their willingness to disclosure personal information so that corporation can focus on fancy marketing campaigns to different individuals. A total of 900 respondents for asking several personal information, concern of privacy, knowledge and direct or indirect experience on Skype.

The information is collected in this survey including: (a) Identification data (e.g. name, cell phone & email address); (b) Profile information (e.g. age, sex, nation & city); (c) information about experience usage and knowledge of the Skype; (d) Rating of information requested regarding to relevance, sensitivity and privacy concerns.

Research Model

Figure 1. showed the conceptualization of the users' intentions to disclosure personal information online. Our research model revels: (1) The effect of users who are willing to provide correct information based on mandatory and voluntary units with or without explain the reason for Skype in collecting information. (2) Behavioural intention to disclose their personal information to the corporation implied by privacy concern, sensitivity and relevant of requested questions. Table 1 is

summarized the users whether they are provided correct information in two treatment methods. We observed the group in Mandatory with explanation are more likely to provide correct information. Alternatively, the group in voluntary without explanation are probably to give incorrect information. Therefore, the effect of trust is play a significant role of information disclosure. Thus, in order to get more correct information from survey, organization should first build trust relationship with users.

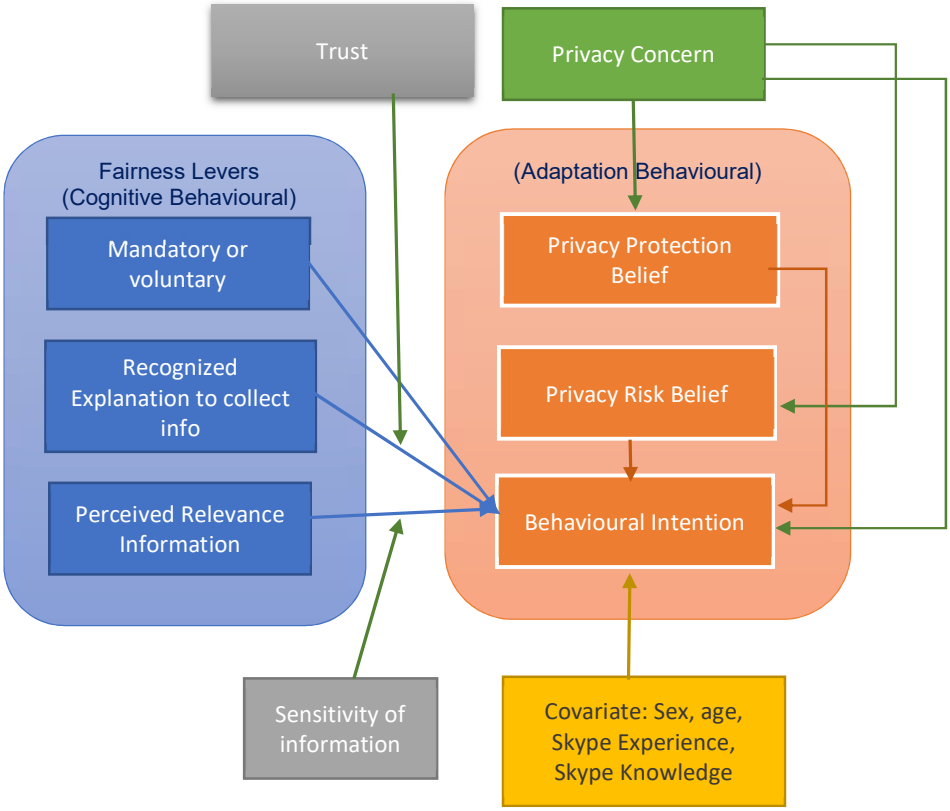


Figure 1. Research Model

Table 1. Categorize that users provided correct information with two treatment methods.

1st treatment	Total no	2nd treatment	Total no.		Total no.
Mandatory	468 (52%)	Provide explanation	243 (52%)	D=1 correct info	205 (84%)
				D=0 incorrect info	38 (16%)
		Do not provide explanation	225 (48%)	D=1 correct info	181 (80%)
				D=0 incorrect info	44 (20%)
Voluntary	432 (48%)	Provide explanation	234 (54%)	D=1 correct info	184 (79%)
				D=0 incorrect info	50 (21%)
		Do not provide explanation	198 (46%)	D=1 correct info	116 (59%)
				D=0 incorrect info	82 (41%)

Fairness levers and Privacy Concern

Three fairness levers are studied with (i) mandatory and voluntary to participant the survey. (ii) provide explanation the reason for Skype in collecting information (iii) Sensitivity and relevance perceived of information requested and their impact on users whether they are willing to provide correction information to the corporation. This finding can determine user feedback regarding to disclosure their personal information.

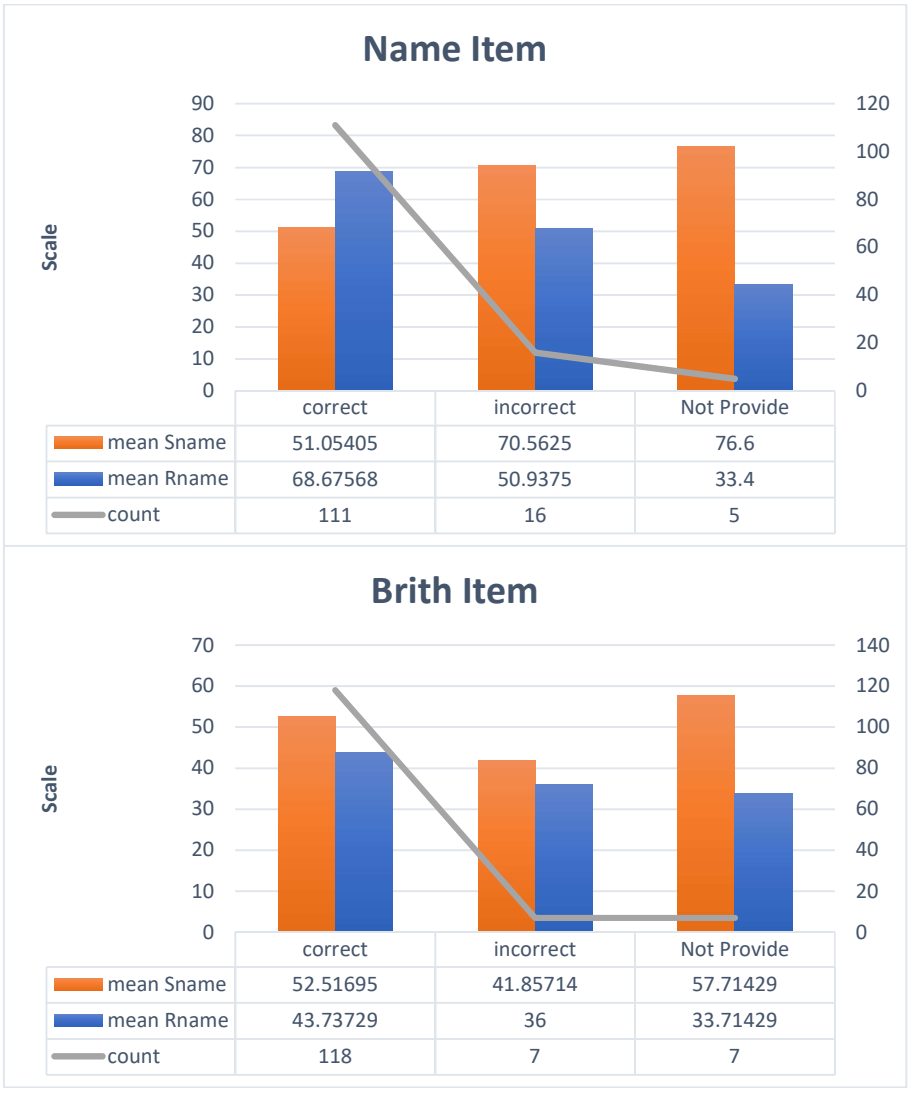
Sensitivity and Relevance of information requested

[2] pointed out that sensitivity has been classified as attribute of personal information. It also reflected the level of affinity to disclosure their information to website. Figure 2. showed distribution of sensitivity and relevance perceived by respondents regarding to specific information requested in the survey. We found that users are less sensitive to provide basic demographic information, like sex, Nation, City and language, while moderately sensitive about email and birth. On the other hand, they are mostly concerned with the name and cell phone number. Table 2. & 3. showed the results of hypothesis testing with several variables on Relevance and Sensitivity.

In addition, we observed that more relevant information requested will enhance users to disclose correct information. Since some individuals may feel suspicion of organization's reliability and honesty, so they do not provide correct information due to distrust. Hence, we revealed that both trust and relevance can enhance the users to disclose correct information.

Hypothesis Testing

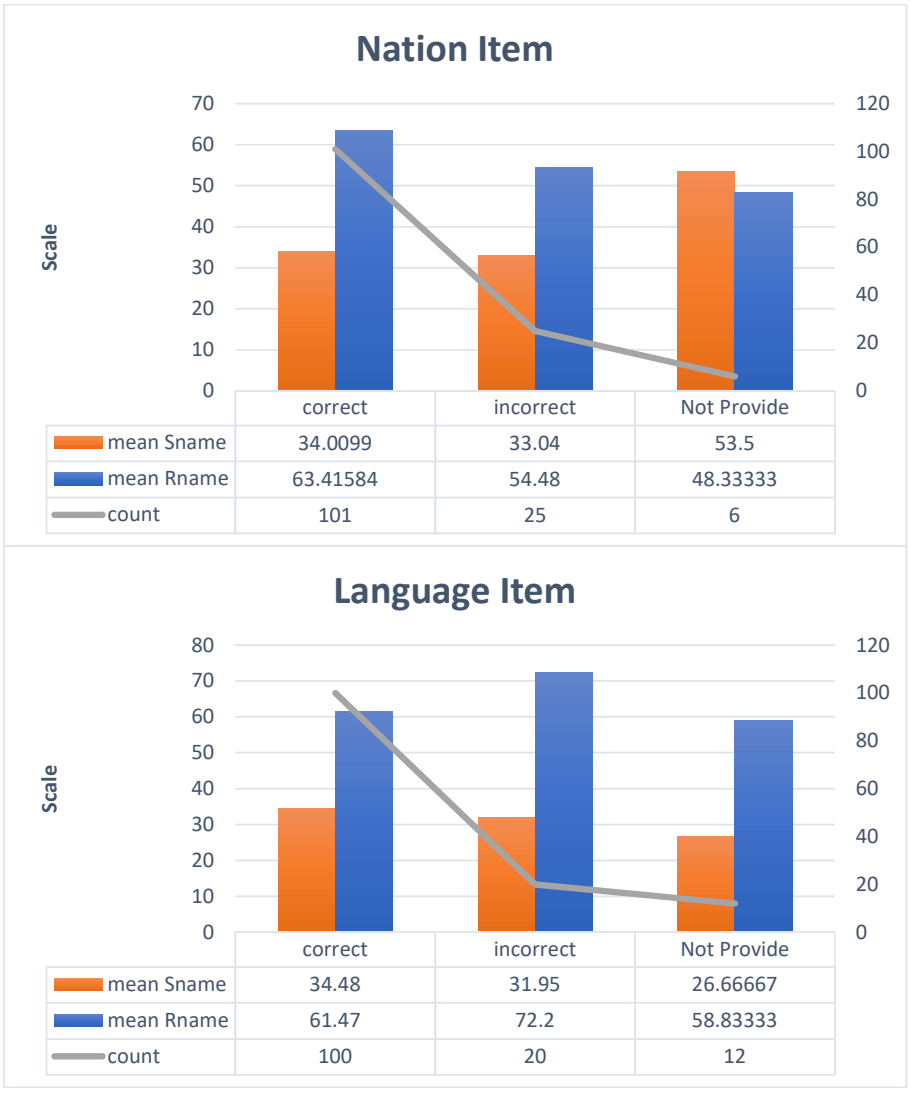
After performing statistical analysis, we decide to remove "not provided" observations for this assessment, because (i) we don't have any information for "not provided observations" (ii) Even removed "not provided" observations, the dataset is still good enough which still maintain high population (n_1=900) with likely no bias compared with original one (n_0=1188).

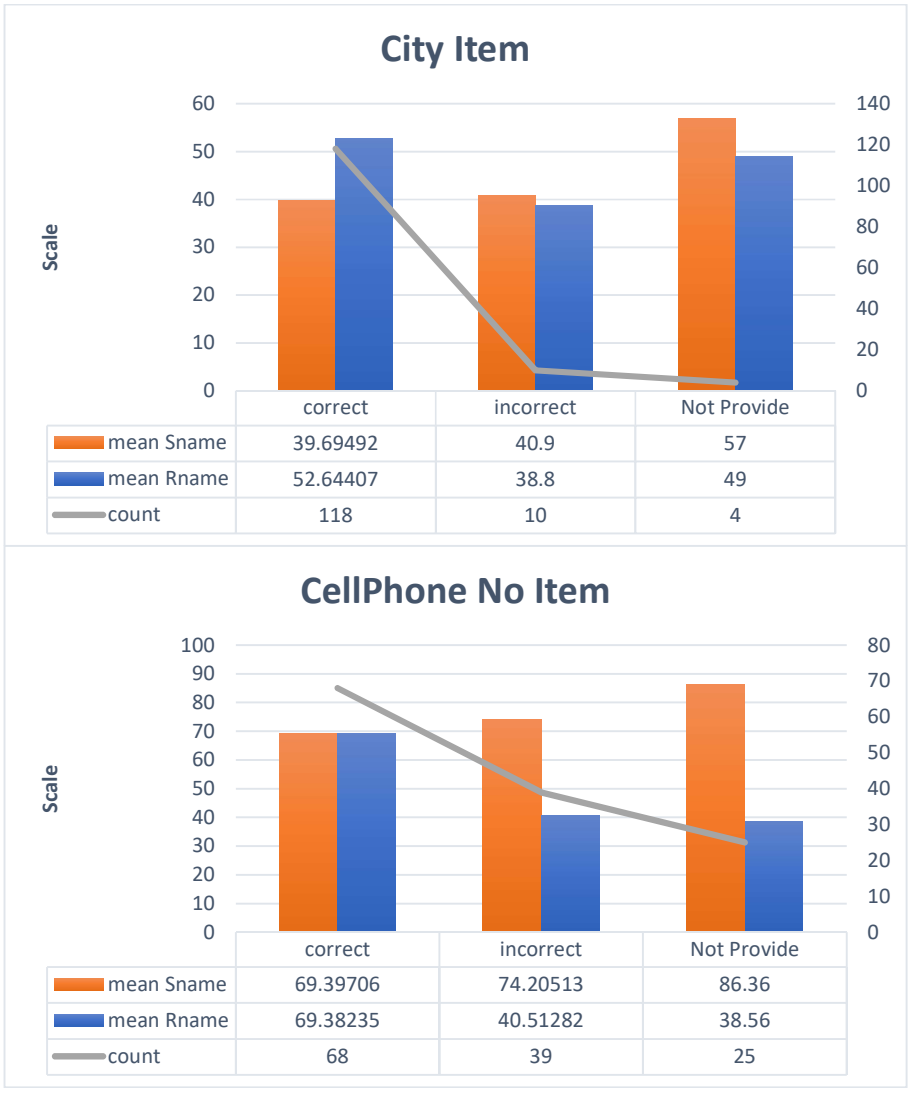

Name Item
Scale
90
80
70
60
50
40
30
20
10
0
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
51.05405
70.5625
76.6
mean Rname
68.67568
50.9375
33.4
count
111
16
5


Email Item
Scale
70
60
50
40
30
20
10
0
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
52.36364
58.66667
65.28571
mean Rname
61.24545
39.2
46.85714
count
110
15
7

Brith Item
Scale
70
60
50
40
30
20
10
0
140
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
52.51695
41.85714
57.71429
mean Rname
43.73729
36
33.71429
count
118
7
7

Gender Item
Scale
60
50
40
30
20
10
0
140
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
36.47009
40.3
41.6
mean Rname
50.11966
38.4
41.6
count
117
10
5


Nation Item
Scale
70
60
50
40
30
20
10
0
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
34.0099
33.04
53.5
mean Rname
63.41584
54.48
48.33333
count
101
25
6


City Item
Scale
60
50
40
30
20
10
0
140
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
39.69492
40.9
57
mean Rname
52.64407
38.8
49
count
118
10
4

Language Item
Scale
80
70
60
50
40
30
20
10
0
120
100
80
60
40
20
0
correct
incorrect
Not Provide
mean Sname
34.48
31.95
26.66667
mean Rname
61.47
72.2
58.83333
count
100
20
12

CellPhone No Item
Scale
100
90
80
70
60
50
40
30
20
10
0
80
70
60
50
40
30
20
10
0
correct
incorrect
Not Provide
mean Sname
69.39706
74.20513
86.36
mean Rname
69.38235
40.51282
38.56
count
68
39
25

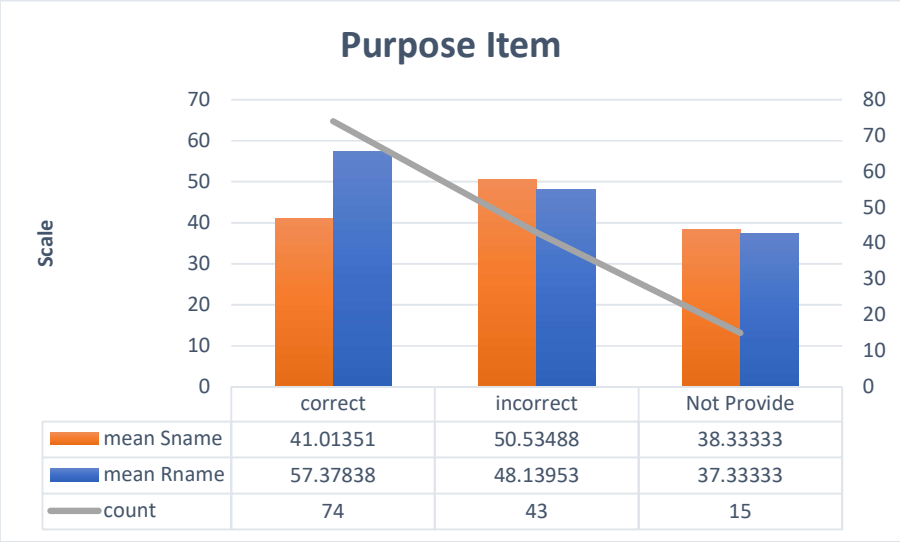


Figure 2 Distribution of sensitivity and relevance on specific information requested in the survey.

Table 2 Summary of hypothesis testing results with other variables on Relevance and Sensitivity.

Variables		Hypothesis	Correlation coefficients	t-statistic	p-value
Name	Relevance	H1: The relevant of information requested has a positive impact on provide correct Name.	2.954025	-3.1799	0.003803 (supported)
	Sensitivity	H2: The negative effect on provide correct Name when sensitive info is requested.	-2.82356	4.0553	0.0002781 (supported)
Email	Relevance	H3: The relevant of information requested has a positive impact on provide correct Email.	2.744275	-3.1493	0.0038 (supported)
	Sensitivity	H4: The negative effect on provide correct Email when sensitive info is requested.	-1.176845	1.3663	0.1821(Not supported)
Birth	Relevance	H5: The relevant of information requested has a positive impact on provide correct Birth.	0.8483692	-1.2952	0.2127 (Not supported)
	Sensitivity	H6: The positive effect on provide correct Birth when sensitive info is requested.	0.2609299	-0.3378	0.74 (Not supported)
Gender	Relevance	H7: The relevant of information requested has a positive impact on provide correct Gender.	1.081194	-1.4022	0.178 (Not supported)
	Sensitivity	H8: The negative effect on provide correct Gender when sensitive info is requested.	-0.4326856	0.79237	0.4367 (Not supported)
Nation	Relevance	H9: The relevant of information requested has a positive impact on provide correct Nation.	1.833391	-1.9175	0.06121 (Not supported)
	Sensitivity	H10: The negative effect on provide correct Nation when sensitive info is requested.	-0.5414064	0.61658	0.5401 (Not supported)
City	Relevance	H11: The relevant of information requested has a positive impact on provide correct City.	1.044182	-1.4037	0.1798 (Not supported)
	Sensitivity	H12: The negative effect on provide correct City when sensitive info is requested.	-0.5545917	0.75048	0.4642 (Not supported)
Language	Relevance	H13: The relevant of information requested has a negative impact on provide correct Language.	-1.058062	1.0649	0.2917 (Not supported)
	Sensitivity	H14: The positive effect on provide correct Language when sensitive info is requested.	0.8348369	-0.9695	0.3362 (Not supported)
Cell phone No.	Relevance	H15: The relevant of information requested has a positive impact on provide correct phone no.	7.457784	-6.5799	1.08e-09 (supported)
	Sensitivity	H16: The relevant of information requested has a negative impact on provide correct phone no.	-2.405043	2.2138	0.02861 (supported)
Purpose	Relevance	H17: The relevant of information requested has a positive impact on provide correct purpose.	2.986815	-2.9697	0.003577 (supported)
	Sensitivity	H18: The relevant of information requested has a negative impact on provide correct purpose.	-1.580037	1.2842	0.2017 (Not supported)

Table 2 Summary of hypothesis testing results.

Variables	Hypothesis	Correlation coefficients	t-statistic	p-value
Explanation	H19: To provide an explanation why Skype needs the item has a positive impact on provide correct information	0.132929	-3.9799	7.497e-05 (supported)
Mandatory	H20: The positive effect on provide correct information when mandatory is requested.	0.1529607	-4.6032	4.806e-06 (supported)
Pre-knowledge	H21: The negative effect on provide correct information when the user has the knowledge on Skype.	-0.007540806	0.22614	0.8211(Not supported)
Experience	H22: The negative effect on provide correct information when the user has the direct or indirect experience of using Skype before.	-0.02594269	0.76419	0.4452 (Not supported)

Pooling Model

Pooling model ignores the unobserved heterogeneity of users, and it also ignores possible association within users (groups). To control for the unobserved heterogeneity, we can take the first difference (period-to-period change) and use it for the analysis by (i) First Difference (FD) Model, or include the dummies indicating each user by (ii) Fixed Effects (FE) Model (or Least Squares Dummy Variable (LSDV) Model) Or use a group(user)-specific random element by (iii) Random Effects (RE) Model.

Difference between the models

FD model is perfectly controls for time-invariant user heterogeneity. While FE Model make the dummy variables indicating each user and apply the regression to the data including all dummy variables. It's perfectly controls for time-invariant user heterogeneity. In the meanwhile, RE model separates individual effects into two pieces: (a) individual effects from observed heterogeneity; (b) individual effects from unobserved heterogeneity, which is a group(user)-specific random element and has a strict assumption that the effects are uncorrelated with the regressors. However, RE model cannot perfectly controls for time-invariant user heterogeneity. Typically, FE model is preferred to RE model, but we can test the observed heterogeneity as it can be incorporated in the RE model. Also, we can save the degree of freedom so we can have a better chance of rejecting the null hypothesis. (See Table 4-5 & Figure 3)

Table 3 Estimate the impact of accuracy for user provides the correct information on the items by OLS.

	OLS						
	OLS			Panel Robust Model		Pooling Model	
Variables	Estimate Coeff.	Std. Error	p-value	Std. Error	p-value	Std. Error	p-value
β_1: constant	0.4312243	0.1393066	0.00203 **	0.1367	0.0017	0.13930660	0.002026 **
β_2: Relevant (R)	0.0031314	0.0031314	7.84e-10 ***	0.0005	<0.0001	0.00050380	7.841e-10 ***
β_3: Sensitive (S)	-0.0015490	0.0004850	0.00145 **	0.0005	0.0019	-0.00154900	0.001453 **
β_4: Explanation (E)	0.1109791	0.0275731	6.18e-05 ***	0.0276	<0.0001	0.11097909	6.184e-05 ***
β_5: Mandatory (M)	0.1196238	0.0273527	1.37e-05 ***	0.0275	<0.0001	0.11962382	1.368e-05 ***
β_6: Pre-knowledge (P)	-0.0095119	0.0347177	0.78416	0.0361	0.7923	-0.00951189	0.784165
β_7: Experience (EX)	-0.0128894	0.0368208	0.72638	0.0354	0.7158	-0.01288942	0.726378
β_8: gender	0.0351895	0.0293783	0.23131	0.0292	0.2281	0.03518953	0.231311
β_9: age	0.0031325	0.0055919	0.57550	0.0053	0.5569	0.00313249	0.575498
F-statistic	11.66					11.665	
p-value	7.265e-16					7.2646e-16	
Observations#1	The R-squared is 0.09481, meaning that the model only explains 9.5 % of the variability of the response data around its mean. Residual standard error is 0.4071. Relevant, Sensitive, Explanation & Mandatory are statistically significant (p-value, shown as Pr(>|t|), < .05). Rather than just comparing the numbers in the output, it can be helpful to visualize the coefficients by coefplot package.			Relevant, Sensitive, Explanation & Mandatory are statistically significant (p-value < .05).		The R-squared is 0.094806, meaning that the model only explains 9.5 % of the variability of the response data around its mean. Relevant, Sensitive, Explanation & Mandatory are statistically significant (p-value < .05).	
Observations#2	Two of variables, "P" and "age" of VIF are >1.6. It means they are slightly correlated with at least one of the other predictors in the model.						
Observations#3	BP test has a p-value 1.084e-08, so we can reject the null hypothesis and infer that heteroscedasticity is indeed present.						

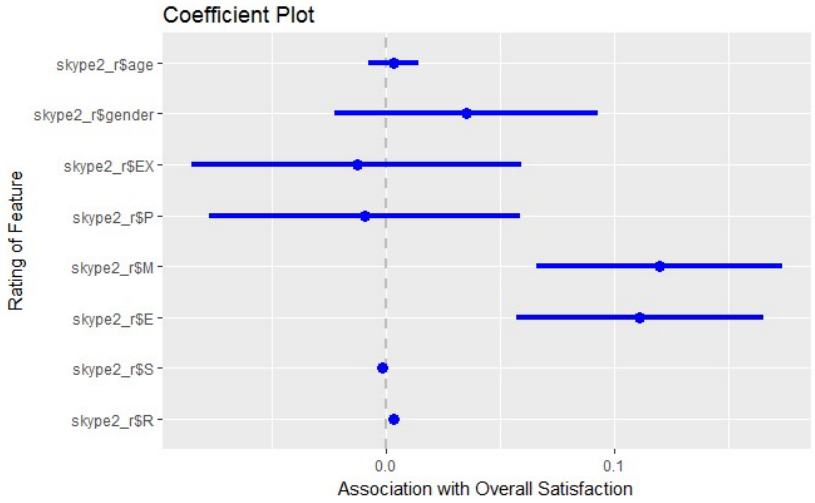


Figure 3. Coefficient Plot of ordinary least squares (OLS).

Correlations

Figure 4. showed the correlation plot, it is shaded blue for positive correlations and red for negative. We found that "correct provide information (I)" is positive correlated with "privacy concern (PC1-PC3)" and " direct or indirect experience of using Skype (EX)".

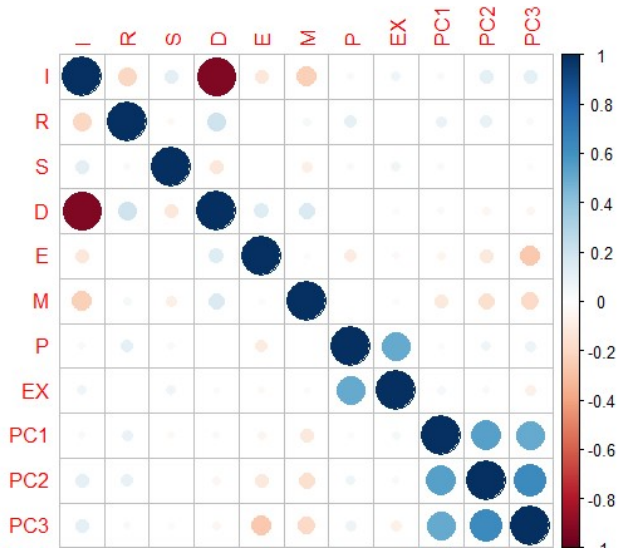


Figure 4 A Correlation plot (red: negative correlation; blue: positive correlation)

Table 4. Estimate the impact of accuracy for user provides the correct information on the items by FD, FE & RE.

$$lnD_{it} = \beta_1 + \beta_2 R_{it} + \beta_3 S_{it} + \beta_4 E_{it} + \beta_5 M_{it} + \beta_6 P_{it} + \beta_7 EX_{it} + \beta_8 gender_i + \beta_9 age_{it} + \varepsilon_{it}$$

	One-way (Individual) Effect First Difference Model (FD)			Fixed Effect Model (FE)			One-way (Individual) Random Effect Model (RE)		
Variables	Est. Coeff.	Std. Error	p-value	Est. Coeff.	Std. Error	p-value	Est. Coeff.	Std. Error	p-value
β_1: constant	-0.03109979	0.01599243	0.0521677				0.45413895	0.25076215	0.07047
β_2: Relevant (R)	0.00197132	0.00052887	0.0002071 ***	0.00189179	0.00050899	0.0002159 ***	0.00223628	0.00048850	5.365e-06 ***
β_3: Sensitive (S)	0.00012037	0.00045738	0.7924831	-0.00072279	0.00048112	0.1334171	-0.00094532	0.00046407	0.04194 *
β_4: Explanation (E)							0.11216390	0.05062956	0.02699 *
β_5: Mandatory (M)							0.12434388	0.05009107	0.01323 *
β_6: Pre-knowledge (P)							-0.00220262	0.06346343	0.97232
β_7: Experience (EX)							-0.01926846	0.06746783	0.77525
β_8: gender							0.03369842	0.05390594	0.53204
β_9: age							0.00288789	0.01026542	0.77853
F-statistic	6.95071			8.26624			4.8287		
p-value	0.0010171			0.0002797			7.7151e-06		
Observations#1	To control for the unobserved heterogeneity, we can take first difference (period-to-period change) and use it for analy			p-value of "S" >0.05, so it is not a significant in the model & just a control variable.			p-value of "P", "EX"," gender" & "age" >0.05, so they are not significant in the model & just control variables.		
Observations#2	Adv: It can remove the latent heterogeneity from the model whether the fixed or random effect model is appropriate.								
Observations#3	Disadv: The differencing also removes any time-invariant variables from the model.								
Findings: (Combined Table 1 & 2)	#1. We found that "P", "EX"," gender" & "age" are not significant in above models. #2. As a general observation, with a variety of approaches available, FD & FE estimators do not have much recommend. #3. Conduct Hausman test to check whether the individual effects are uncorrelated with the regressors, we found that p-value is 0.008902 which is significant, therefore we use Fixed Effect Model (FE).								

Table 5. Panel glm model for Maximum Likelihood estimation.

Variables	Estimate	Std error	p-value	Findings
β_1: constant	-0.684385	2.07591	0.7416	p-value of "R" "S", "E" &"M" <0.05, so they are significant in the glm model.
β_2: Relevant (R)	0.017898	0.00415	1.58e-05 ***	
β_3: Sensitive (S)	-0.007858	0.00391	0.0445 *	
β_4: Explanation (E)	0.854014	0.41465	0.0394 *	
β_5: Mandatory (M)	0.809647	0.40778	0.0471 *	
β_6: Pre-knowledge (P)	-0.187147	0.51546	0.7166	
β_7: Experience (EX)	-0.158596	0.54998	0.7731	
β_8: gender	0.183853	0.44289	0.6781	
β_9: age	0.039645	0.08567	0.6435	
sigma	2.362018	0.32358	2.88e-13 ***	

Moderation effect

To test the moderation effect, we include the interaction term between the predictor and the moderator, and test the coefficient of the interaction as shown in Table 6.

$$R = c + \beta_1 S + \beta_2 D + \beta_3 S * D + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon$$

Then, we have two models for fitting the data.

$$R_1 = c + \beta_1 S + \beta_2 + \beta_3 S + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 1$$

$$R_1 = c + (\beta_1 + \beta_{3)} S + \beta_2 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 1$$

$$R_0 = c + \beta_1 S + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 0$$

Compare the fit (R-square) between the restricted model and the unrestricted model

$$R = c + \beta_1 S + \beta_2 D + \beta_3 S * D + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad unrestricted\ model$$

$$R = c + \beta_1 S + \beta_2 D + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad restricted\ model\ by\ \beta_3 = 0$$

Table 6. Main effect and interaction effect on R, S and D.

	F-statistic	p-value	Findings	Wald test
Main effect model	20.71	1.612e-09	There is main effect model of relevant on the relationship between user provides the correct information on the items.	The purpose of Wald test is to find out if explanatory variables in the models are significant. We found that Chi-squared X2 is 11.4 & P(> X2) 0.00075. It shows the parameters are not zero, which means we should include variables in the Moderation effect model.
	R=46.48476 -0.01005 S +13.31449 D			
Moderation effect model	13.9	7.299e-09	There is moderation effect model of relevant on the relationship between user provides the correct information on the items.	
	$R =$ 48.11815 − 0.03989 S + 11.19904 D + 0.04015 S*D			

Further study and Suggestions for (i) screening &(ii) main effect and interaction effect on R, S and D by ANOVA as shown in Table 7: From results of ANOVA, both result of Pr(>F) on relationship between user provides the correct information on the items "D" is very close to zero, which mean "D" is a significant predictor, while the interaction of R, S and D is less significant. Furthermore, we performed model comparison in AVOVA. Two models are found significant, namely (i) Model 2 which is the significant on R, S and D; (ii) Model 5 which is the significant on R, S, D, E, M & P. We do better to have more data to establish the relationship between D and S, R. The possible control variables like Nation, City and preferred language can incorporate into the model.

Conclusions

The effect of trust and relevant information requested are play important roles which intended users to disclose correct information.

Corporation can make use this findings and better understanding of customers' value to strengthen Skype performance and create competitive advantage in several aspects.

References

[1] Zimmer, Christopher J; Arsal, Riza Ergun; Marzouq, Mohammad Al; Grover, Varun;, "Investigating online information disclosure: Effect of Information revelance, trust and risk," *Information & Management,* pp. 115-124, 2010.

[2] Angst, C. M; Agarwal, R.;, "Adoption of electronic health records in the presence of privacy concerns: the elaboration likeihood model and individual persuasion.," *MIS Quarterly,* vol. 33, p. 2, 2009.