DA2 Term Project - Survey Analysis - Cosmin Ticu

Executive Summary

The aim of this study is to uncover the patterns of association between how office employees perceive their engagement in corporate safety missions and how they perceive various office safety factors using a probability model. With survey data stemming from a Fortune 100 company and a large (5000+) pool of respondents, training and test samples are used to validate findings. While the data and subsequent models are not fit for a classification task due to a heavily skewed response rate, due to a hypothesized few influential respondents, the study was able to draw minor predictive findings. Most notable of findings was that office employees who have not reviewed their workplace's hazards, who report not feeling safe in the office and who report not being familiar with office risks are less likely, on average, to feel engaged to a corporate safety mission. The findings of this study can be made more robust by minimizing the non-classical measurement error attributed to reactive "I agree" answers and by minimizing the usage of observational variables, rather moving towards latent questions.

Introduction - Purpose & Rationale

The purpose of this study is to inspect the pattern of association between employees' perception of safety factors within their workplaces and how engaged they feel to their respective corporate safety missions, which ideally serve the purpose of protecting them and providing regulations for safe behavior. This safety is understood as a factor of incidents. Understood broadly, it refers to factors such as knowledge of fire escapes, injuries in the workplace, employee group behavior, knowledge of what to do during a crisis situation, etc. With an identified pattern of association, the ultimate goal of the study is to build a predictive statistical model which can allow us to estimate the trust or engagement of an employee into the safety measures preached by their employer. This can be extremely useful for companies that want to make their workplaces feel safer to their employees. The research question can be summed up as: What are the factors that impact employees' engagement in corporate safety missions and how can we predict whether an employee will be engaged or not?

Data collection

With the broad research idea in mind, the data for this study comes from a Fortune 100 global company in the form of a safety survey conducted on over 5000 employees in different working environments (from offices to manufacturing and services). The original survey contains answers to questions pertaining to employee safety within all the working environments (the original data with questions can be found here). For the sake of narrowing down the scope of this analysis and based on the client's needs, the data was filtered (R script for cleaning can be found here) to only contain the office environment questions. As such, the research question is narrowed down to office workplace safety. Nonetheless, for future analysis and robustness checks, the other environments can be employed. It is, however, beyond the scope of this study.

The sample can be hypothesized to be representative in the case of the office environment, as we are dealing with over 1800 observations, each representing a single employee. The survey was anonymously distributed throughout the company, any employee within the office environment having access to the survey. As such, the sampling distribution does not pose a hiccup in proceeding with generalizing findings. However, measurement errors do.

As with any survey data, classical measurement error is a concern for analysis. The survey was distributed over a longer period of time, without controlling for the environment that the employees were actually in

while answering, while also allowing for employees to leave answers as blank or accidentally skip certain questions and have them marked as answered. As with classical measurement error, it is 0 on average, but it is assumed to be present within all the variables (thus both explanatory and dependent). That is especially prominent because most of the questions are observation-based, attempting to extract opinionated and biased answers from employees which are thought of as theoretically representing the truth. Ideally, this means that we can expect a less steep slope due to classical error in the explanatory variable(s) and a much larger standard error (or spread) in the predicted values and coefficients due to classical error in the dependent variable.

Lastly, non-classical measurement error is also hypothesized to be present within the data, as employees are inclined to always agree with any mission statements provided by their corporate employers. While the survey was anonymous, it is possible that people gave reactive answers, agreeing with the general mission statements and office attitudes. This type of error is much harder to identify and control, and it is suspected to be present and affect the distribution of predicted values (by skewing towards "yes" irrespective of explanatory parameters).

The data

The variables of choice within the office environment sample are safe_office, hazards_reviewed, office_familiar_risks and safety_mission engagement.

Figure 1: Distribution of safety mission engagement answers

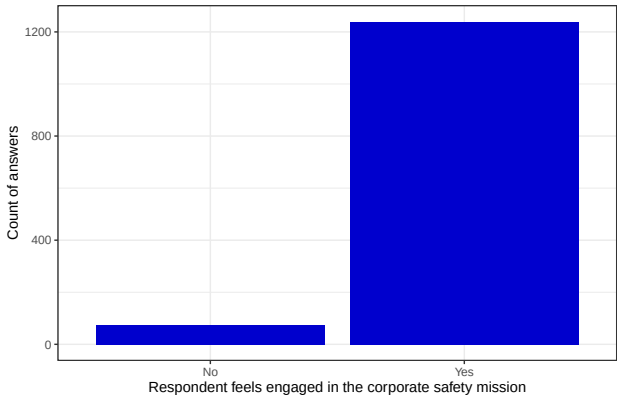


Figure 2: Distribution of office safety answers

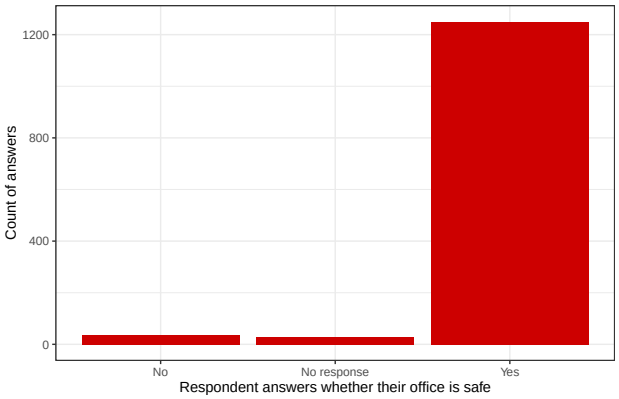


Figure 3: Distribution of office risks familiarity answers

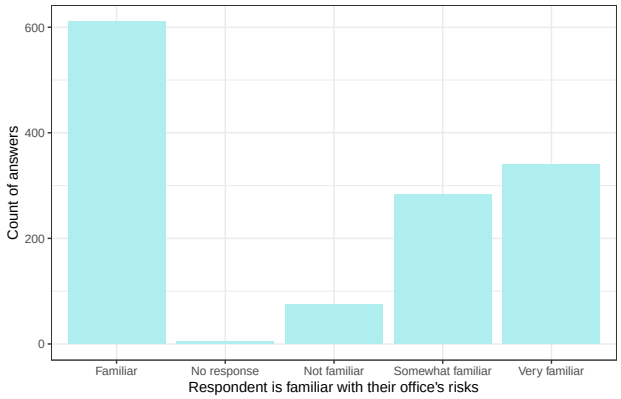
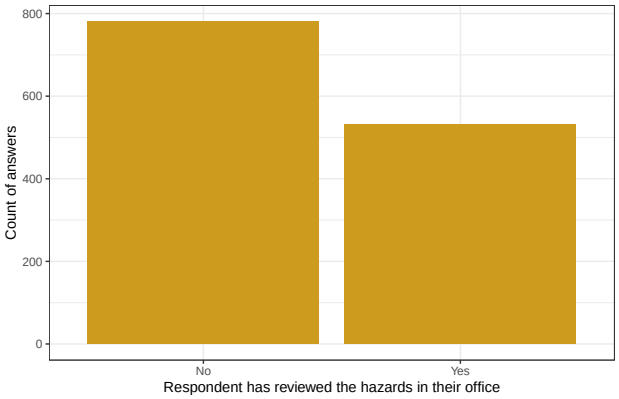


Figure 4: Distribution of office hazard review answers



As per the research question, the safety_mission_engagement is chosen as the dependent variable, as it is of most interest to the client of this study. This study only employs binary categorical variables in saturated probability models. All the variables are transformed into binary and none of the observations are dropped, even if there are "no response" occurrences within the data. That is because the outcome variable only contains "yes" and "no" answers, thus making even a lack of response in one of the explanatory variables as potentially useful for uncovering the full pattern of association. Furthermore, with an almost 95% "yes"

response to the safety mission engagement, it is paramount to find out just how influential those 5% of employees are who do not feel engaged. This will be uncovered by statistically significant coefficients.

The Likert scale variable of office risk familiarity is transformed into multiple binary variables (baseline kept at "Familiar"). This is done instead of giving each entry a numeric value in order to reduce potential bias (because the difference between "Very familiar" and "Familiar" might not be quantitatively the same as between "Somewhat familiar" and "Not familiar"). Further summary statistics on each of the created binary variables can be found within Appendix A.1.

What we can generally see, however, is that only the hazards_reviewed variable has an even distribution between "yes" and "no" answers. This variable is hypothesized to suffer the least from classical measurement error and not be impacted by the non-classical error, because it is a latent variable (see its associated question in original survey).

With the variables of choice and the probability model framework in mind, we can proceed with creating the linear probability models. For this, training (75% or 1300 observations) and test (25% or 500 observations) samples have been created in order to benchmark the performance of the final model of choice against fresh data as a robustness check.

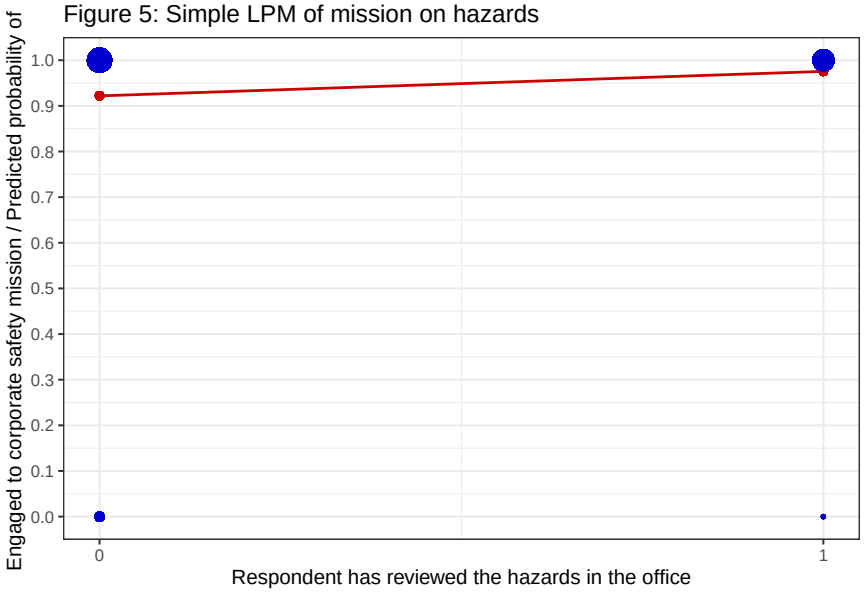
Analysis and modelling

This section provides a walkthrough of the linear probability models (LPM) (a coefficient comparison of which is available here) and the logit and probit models used for actual prediction. First, an LPM rich model is built on which inferences are drawn. Then, it is used with bounded non-linear probability models to assess prediction and coverage (Pseudo R-Squared model summaries available here and model summaries with marginal differences for interpretation available here).

LPM

The simple LPM is built by regressing safety_mission_engagement on hazards_reviewed. As can be seen in the model summary HTML, the simple LPM has a statistically significant slope coefficient. However, what the simple linear probability model does is that because 61 respondents who have not reviewed the office hazards also do not feel engaged in the safety mission, it just classifies all employees that have not reviewed the office hazards as also not being engaged in the safety mission. This can be clearly seen in Appendix A.2, comparing between A.2.1 and A.2.2. While this is a naive model, it sets a statistically significant basis on which to create a richer model.

Figure 5: Simple LPM of mission on hazards

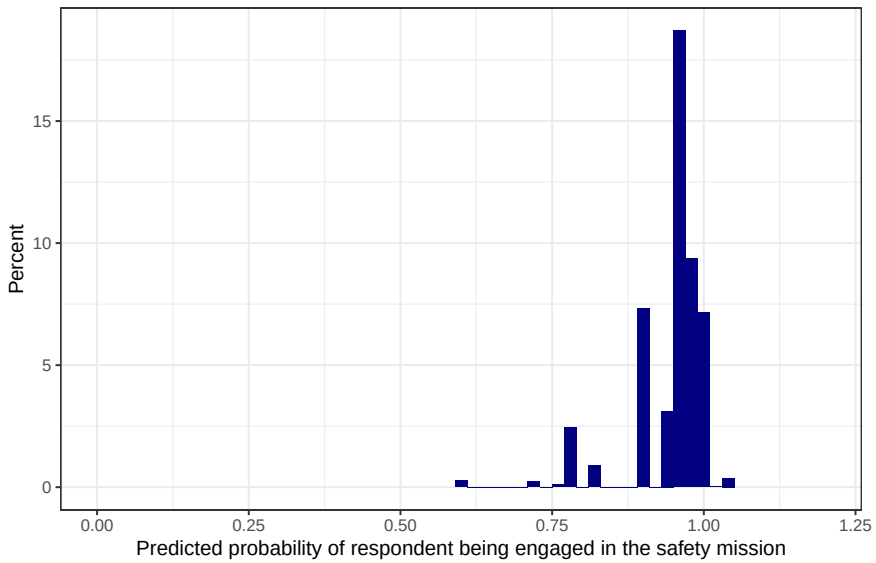


The next 3 LPM models (found in the analysis R script) each add a new explantory variable used with the as.factor() function R to create dummy variables. The richest model, the 4th one, takes the form of regressing safety_mission_engagement on safe_office, hazards_reviewed and office_familiar_risks, thus containing all the variables of choice. This one also has the largest R-squared, which, even though not interpretable, means that it becomes the model of choice.

Model of choice - rich LPM

Because we have an upper boundary that is larger than 1 for these values, we need to use a probit or a logit model for prediction. However, for the purpose of uncovering patterns of association, we can proceed with using this rich LPM. Figure 6 below showcases the predicted probability distribution of our rich LPM model which we can use to inspect the bottom 10% of respondents and the top 10% of respondents to identify some key characteristics and differences.

Figure 6: Predicted probability distribution of LPM rich



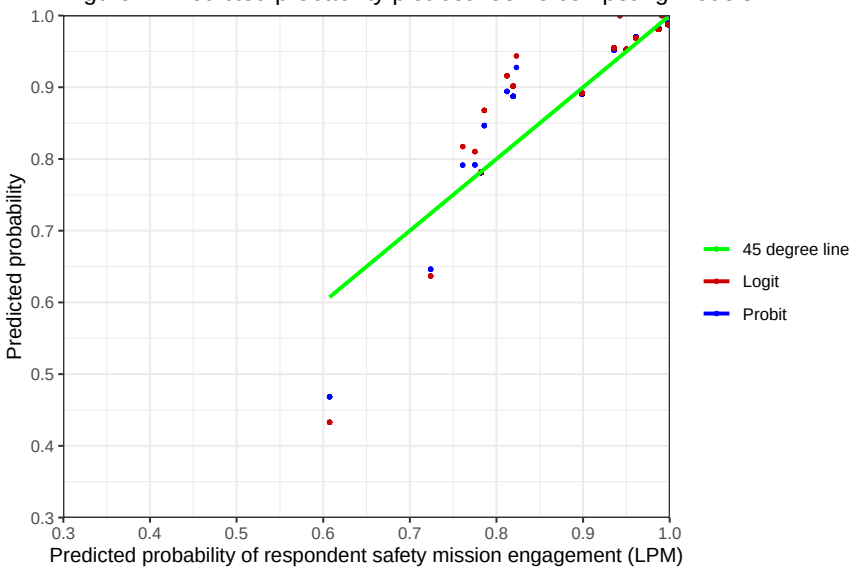
The following interpretation is based on the summary statistics tables in Appendices A.3.2 to A.3.7. The bottom 10% means lower probability (starting from 60%) of safety mission engagement, while the top 10% of predicted values refers to a very high (also above 100% in the case of this unsaturated model) probability of safety mission engagement. A few interesting findings are that most of the people that reported their office as not safe are in the bottom 10% of predicted safety mission engagement, while in the top 10% we have the majority of respondents that did not answer the office safety question. There are only employees that are familiar with office risks in the top 10% of predicted safety mission engagement. It also appears that the hazard variable has a similar distribution for both the top 10% and the bottom 10% of predicted safety mission engagement, even though its coefficient is statistically significant in all the LPM models. These are useful findings because they pave the way for building probit and logit models in hopes of findings statistically significant coefficients (like the LPM models) but which are also viable for predictive analytics.

Logit and probit models

The logit and probit models were run on the same formula used by the rich LPM and their predicted probabilities are shown with summary statistics in appendix A.4.1. We can clearly see there that the non-linear models adjusted the upper boundary to 100%, while reducing the lower boundary to around 45% percent (43% logit and 47% probit). This gives us a larger spread of predicted values (but not a more normal distribution, as can be seen in appendices A.4.4 and A.4.5), which might attenuate some of the effects caused by the biased reactive "yes" answers in the survey or might showcase that the non-linear models tend to underestimate smaller probabilities and overestimate larger probabilities. This needs to be visualized in a

comparison of predicted probabilities on a plot against the linear model.

Figure 7: Predicted probability plot between 3 competing models



What we can see from Figure 7 is that the non-linear models appear to be underestimating the lower probabilities in comparison to the LPM rich model, while overestimating the higher probabilities. The only way to check if the models are suitable for prediction is to run a bias analysis and a calibration curve.

From the model comparison with computed Pseudo R-squared we can see that models 2 (logit) and 3 (probit) are very similar in terms of Pseudo R-squared, AIC and BIC, but very different in terms of coefficients. However, these coefficients cannot be interpreted by themselves, but rather they need to be converted into marginal differences. The comparison between models with marginal differences can be found here.

From this point on, because of the similarity between the two non-linear models, the logit is arbitrarily chosen as the model of choice. Before interpreting its coefficients, we need to inspect its goodness of fit with regards to the earliest naive model, in order to see if the dispersion (distribution) of fitted values shows a clearer distinction between the respondents that actually said they feel engaged in the safety mission and the ones that do not feel that way.

Figure 8: Distribution of simple LPM predicted probabilities by outcome

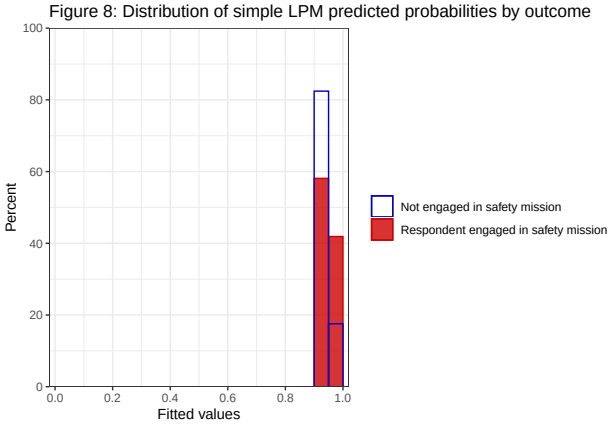
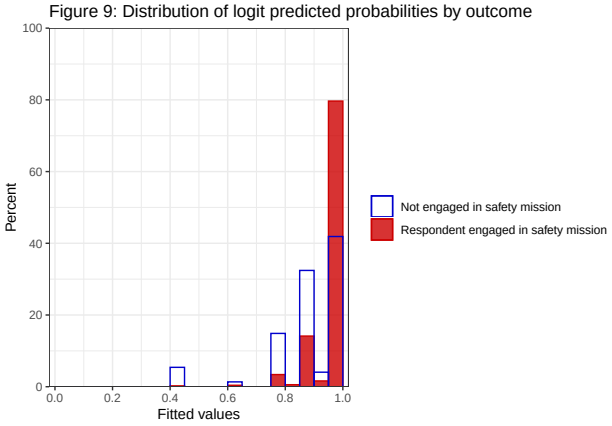


Figure 9: Distribution of logit predicted probabilities by outcome



Figures 8 and 9 show a clear difference between the naive linear probability model and the logit model. We can see that the distribution is wider between actual "no" outcomes and "yes" outcomes within the logit model's predicted values. However, there is still noticeable overlap between the two. This tendency towards the "yes" outcome, perhaps even due to the non-classical measurement error, makes this model unfit for a classification task, as the logical 50% boundary would result in a mere few instances being categorized

accurately for the "no" outcome, much less than the actual number. From the looks of this model, it would have a tendency to commit few false negatives at the expense of committing many false positives. It makes the statistically significant coefficients all the more interesting to inspect. For future analysis purposes, it would be interesting to investigate whether there are a few influential observations, which might actually in turn enable the statistically significant but fairly small slope coefficients. Lastly, as can be observed in the summary statistics on conditional predicted probabilities in appendices A.4.2 and A.4.3, the predicted probability means between the two outcome groups ("yes" table versus "no" table) are extremely close to each other.

Bias & calibration curve

With a bias around 0.0000000006, we can safely dismiss concerns about a biased predicted sample. Furthermore, as can be seen in appendix A.5.1, the calibration curve (even with 15 bins) shows probabilities scattered around their actual ones, with a few influential or more scattered points showing a few abrupt patterns. Nonetheless, the logit model's calibration curve is close enough to the 45-degree line in order to warrant this model a somewhat calibrated mark. This means that we can carefully proceed with the final model interpretation.

Final model & Interpretation

Logit regression on LPM rich formula: **mission = link_function(1.452 + 0.937 * hazards + 14.384 * no_response_familiar_risks - 1.722 * not_familiar_risks - 0.891 * somewhat_familiar_risks + 0.429 * very_familiar_risks + 16.041 * no_respose_safe_office + 1.548 * yes_safe_office)**

Based on the marginal differences computed here, we can proceed to draw interesting findings from the interpretation of only the statistically significant coefficients.

As such, office employees in the context of this study that answer that they believe their office is safe are 13.3% more likely, on average, to answer that they feel engaged in the corporate safety mission, if we control for their answers on whether the office hazards have been reviewed and for their answers on whether they are familiar with the office risks. Another finding would be that office employees in the context of this study that say they are not familiar with their office's risks, that say they are not safe in the office and that say they have not reviewed their office's hazards are 15% less likely to answer that they feel engaged in the corporate safety mission. These findings show that, while the model is not suitable for a classification task due to the heavy bias towards the "yes" answers, the few possibly influential observations (as respondents who do not feel engaged to the corporate safety mission) have uncovered some interesting patterns of association between perceived office environment safety and perceived engagement to a corporate mission.

Robustness check

As a final step of this study, a series of robustness checks was conducted on the test sample of the data that was created as a 25% representative sample of the original office data. The summary statistics (appendix A.6.1) of the test sample predicted values appear very similar to their counterparts in the training sample (appendix A.4.1). The calibration curve of the logit model run on the test data (appendix A.6.2) shows less calibration, while the comparison between the distribution of logit predicted probabilities by outcome on each of the samples shows a distribution on the test sample much more similar to the original simple LPM naive model, rather than the richer logit model. As a last robustness check, it appears (in this HTML model comparison document with marginal differences) that, while not all of the coefficients have retained their significance level, most of the test sample coefficients are within 2 standard errors away from their counterparts in the training sample. Lastly, it appears (in this HTML model comparison document with the actual coefficients) that the Pseudo R-squared takes a large dip in test sample. On a concluding remark, perhaps taking a 25% test sample of a sample already 95%-dominated by "yes" mission statement engagement answers has made the few influential points too far and few.

Appendix A

This appendix serves the purpose of storing and displaying tables, charts, graphs and model comparisons, all of which supplement the main body of this text. References to the appendix can be found throughout the study.

A.1: Descriptive Statistics

Table 1: A.1.1: Descriptive statistics of the chosen variables in the office-side survey

statistics	yes_safe_office	no_safe_office	noresponse_safe_office	hazards	mission
mean	0.9520183	0.0266565	0.0213252	0.4051790	0.9436405
median	1.0000000	0.0000000	0.0000000	0.0000000	1.0000000
min	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
max	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
sd	0.2138090	0.1611388	0.1445211	0.4911137	0.2307025
# missing	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
# used obs	1313.0000000	1313.0000000	1313.0000000	1313.0000000	1313.0000000

Table 2: A.1.2: Frequency distribution of the office risks familiarity variable (Likert scale)

Var1	Freq
Familiar	611
No response	4
Not familiar	75
Somewhat familiar	283
Very familiar	340

A.2: Simple LPM Model

Table 3: A.2.1: Confusion table showing distribution of actual outcome (rows) and explanatory (columns)

	0	1
0	61	13
1	720	519

Table 4: A.2.2: Confusion table showing distribution of predicted outcome (rows) and explanatory (columns)

	0	1
0.92189500640208	781	0
0.975563909774466	0	532

A.3: Model of choice - rich LPM

Table 5: A.3.1: Summary statistics LPM rich predicted values

statistics	pred_lpm_4
mean	0.9436405
median	0.9500083
sd	0.0602673
min	0.6073392
max	1.0415219
# missing	0.0000000
# used obs	1313.0000000

Table 6: A.3.2: Summary statistics of hazards variable on the bottom 10% of respondents

statistics	hazards
mean	0.4060150
median	0.0000000
sd	0.4929441

Table 7: A.3.3: Summary statistics of office_familiar_risks variable on the bottom 10% of respondents

Var1	Freq
Familiar	10
Not familiar	65
Somewhat familiar	52
Very familiar	6

Table 8: A.3.4: Summary statistics of safe_office variable on the bottom 10% of respondents

Var1	Freq
No	32
Yes	101

Table 9: A.3.5: Summary statistics of hazards variable on the top 10% of respondents

statistics	hazards
mean	0.4000000
median	0.0000000
sd	0.4917931

Table 10: A.3.6: Summary statistics of office_familiar_risks variable on the top 10% of respondents

Var1	Freq
Familiar	20
No response	4
Very familiar	106

Table 11: A.3.7: Summary statistics of safe_office variable on the top 10% of respondents

Var1	Freq
No response	23
Yes	107

A.4 Probit & Logit models

Table 12: A.4.1: Summary statistics for the predicted values of the LPM, logit and probit models of choice

statistics	pred_logit	pred_probit	pred_lpm_4
mean	0.9436405	0.9436547	0.9436405
median	0.9546163	0.9523332	0.9500083
sd	0.0640816	0.0636160	0.0602673
min	0.4327784	0.4683971	0.6073392
max	1.0000000	1.0000000	1.0415219
# missing	0.0000000	0.0000000	0.0000000
# used obs	1313.0000000	1313.0000000	1313.0000000

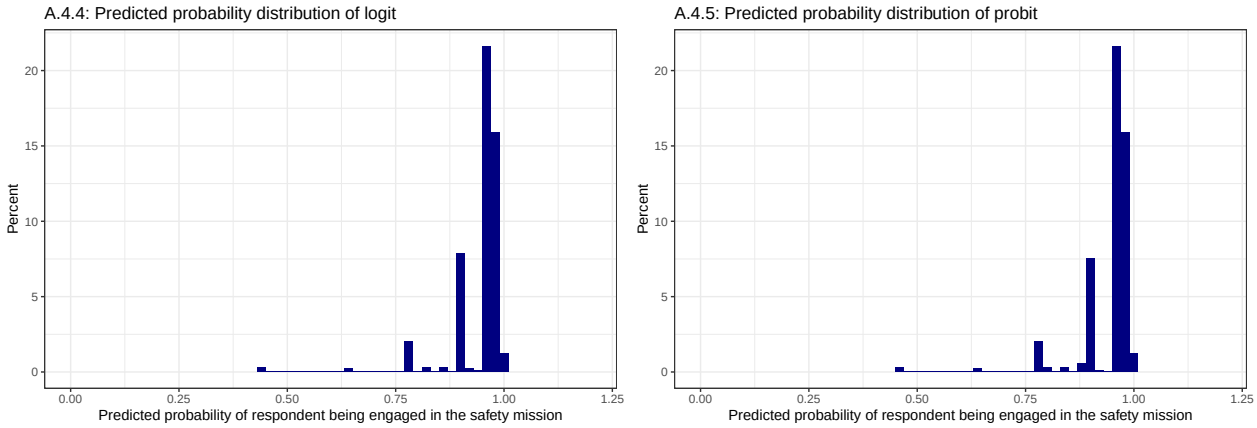
Table 13: A.4.2: Outcome conditional ('yes') summary statistics for the predicted values of models of choice

statistics	pred_lpm_4	pred_logit	pred_probit
mean	0.9474867	0.9477300	0.9477439
median	0.9500083	0.9546163	0.9523332
min	0.6073392	0.4327784	0.4683971
max	1.0415219	1.0000000	1.0000000
sd	0.0552690	0.0557771	0.0561122

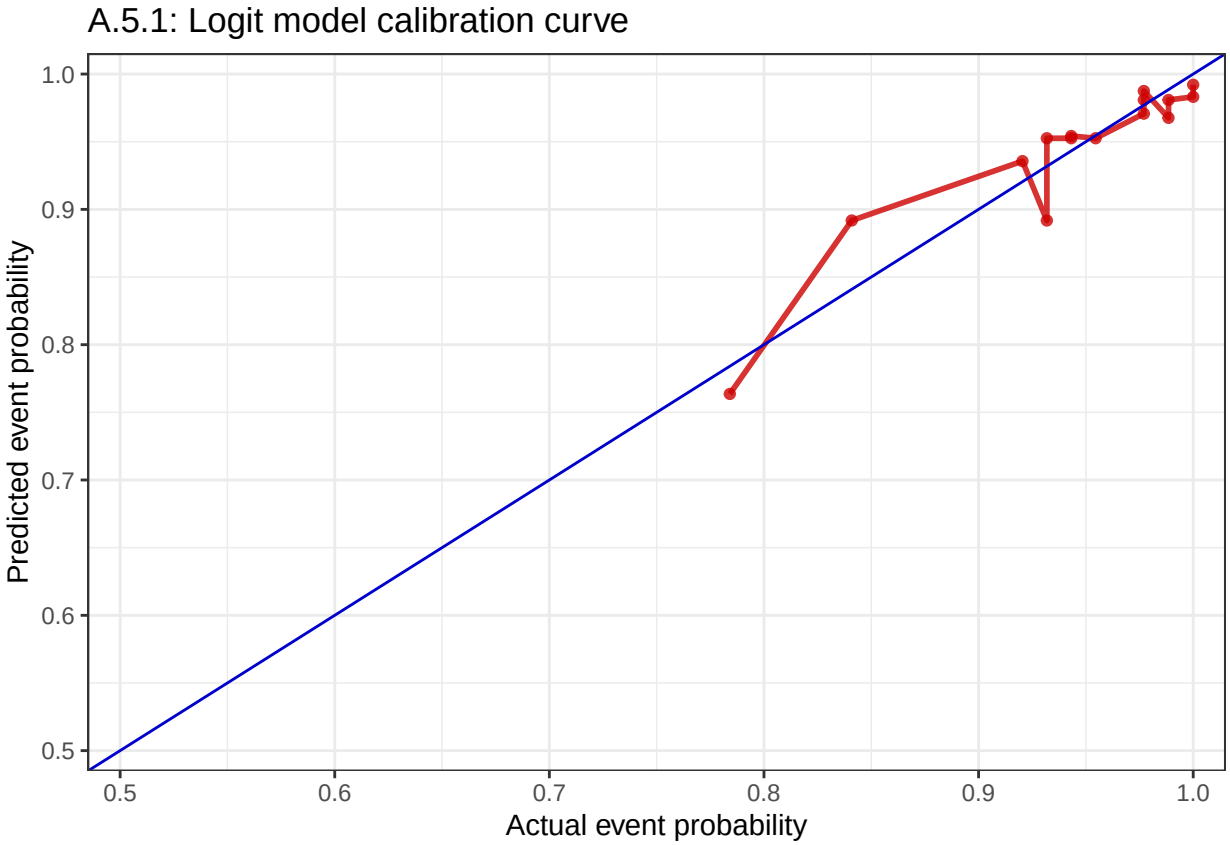
Table 14: A.4.3: Outcome conditional ('no') summary statistics for the predicted values of models of choice

statistics	pred_lpm_4	pred_logit	pred_probit
mean	0.8792436	0.8751691	0.8751878
median	0.8989190	0.8917837	0.8907058
min	0.6073392	0.4327784	0.4683971
max	0.9979760	0.9874586	0.9896511

statistics	pred_lpm_4	pred_logit	pred_probit
sd	0.0949760	0.1265175	0.1195948



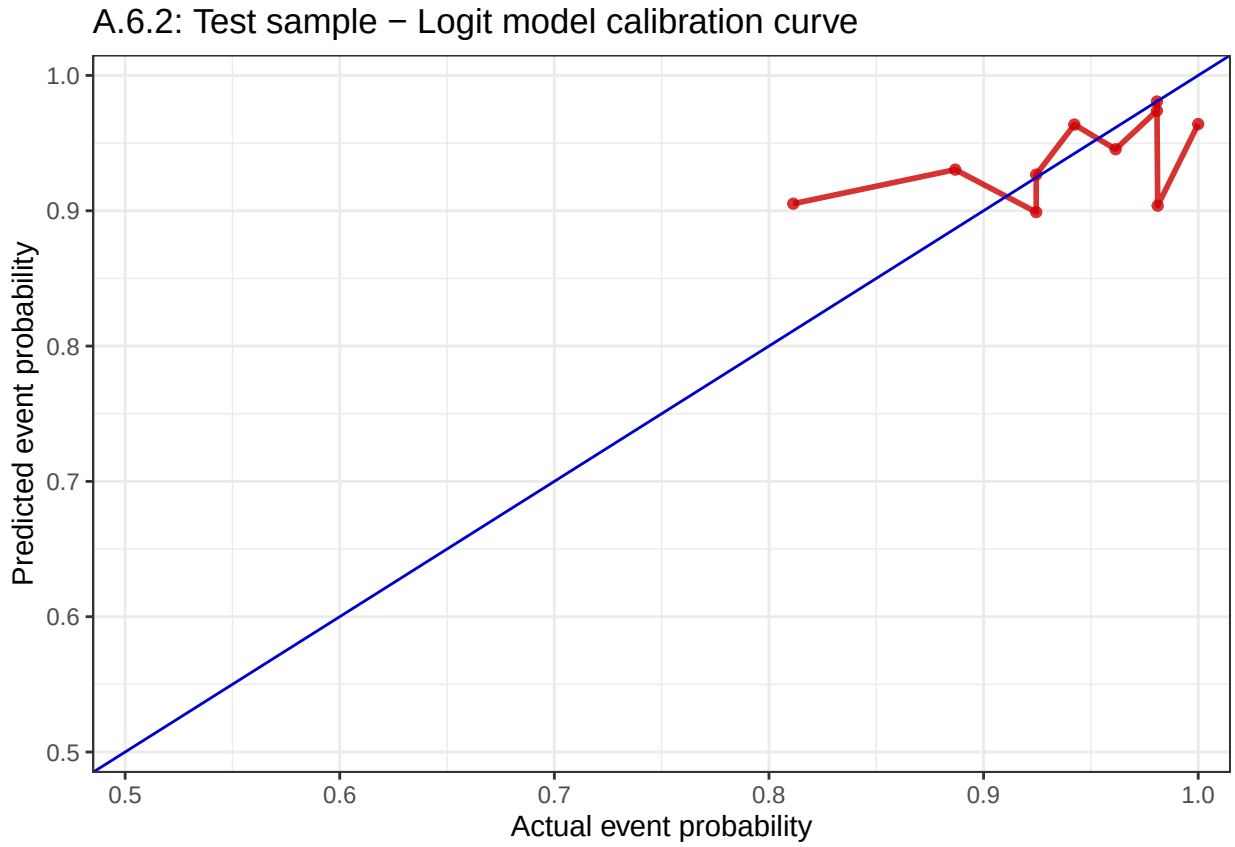
A.5 Bias & calibration curve

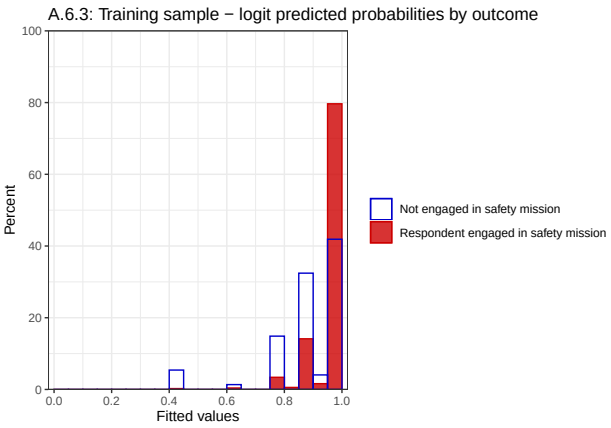


A.6 Robustness check

Table 15: A.6.1: Test sample - Summary statistics logit predicted values

statistics	pred_logit
mean	0.9390476
median	0.9304016
sd	0.0301478
min	0.8565319
max	1.0000000
# missing	0.0000000
# used obs	525.0000000





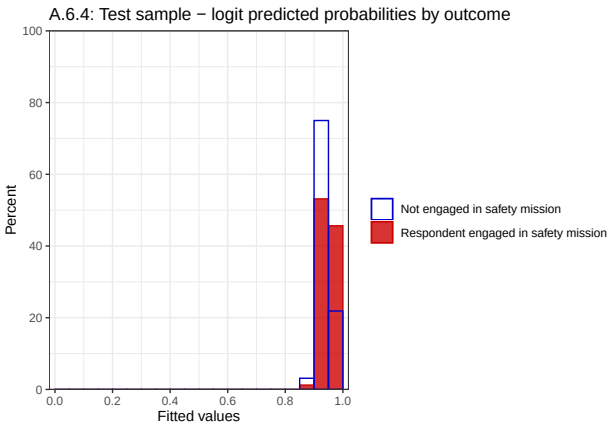


Table 16: A.6.5: Test sample - Outcome conditional ('yes') summary statistics for the predicted values of models of choice

statistics	pred_logit
mean	0.9400115
median	0.9304016
min	0.8565319
max	1.0000000
sd	0.0300865

Table 17: A.6.6: Test sample - Outcome conditional ('no') summary statistics for the predicted values of models of choice

statistics	pred_logit
mean	0.9241973
median	0.9222927
min	0.8565319
max	0.9743226
sd	0.0274661