

**PREVENTING SYSTEMS ENGINEERING FAILURES WITH
CROWDSOURCING: INSTRUCTOR RECOMMENDATIONS AND
STUDENT FEEDBACK IN PROJECT-BASED LEARNING**

by

Georgios Georgalis

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



School of Aeronautics and Astronautics

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Karen Marais, Chair

School of Aeronautics and Astronautics

Dr. Steven Landry

School of Aeronautics and Astronautics

Dr. Bruno Ribeiro

Department of Computer Science

Dr. Vinayak Rao

Department of Statistics

Approved by:

Dr. Gregory A. Blaisdell

To my mother

“There are no goodbyes for us. Wherever you are, you will always be in my heart.”

ACKNOWLEDGMENTS

My doctoral journey has been full of challenging and rewarding experiences, which would not have been the same without the people that spent time with me in professional or friendly settings. It is with great pride and joy that I acknowledge you all here, for the good and bad moments we went through together.

My time in graduate school would not have been so special without the guidance of my advisor, Dr. Karen Marais. She has been one of the most passionate individuals about teaching and research I have ever come across, and she undoubtedly inspired me to be who I am today. I also want to thank my committee members who have been very helpful through their various contributions to this dissertation. Dr. Steven Landry, thank you for agreeing to be in my committee and nudging me to do proper statistics. Dr. Vinayak Rao, thank you for taking a chance on advising me, even if I did not have a statistics background. Dr. Bruno Ribeiro, thank you for being so approachable, I have enjoyed all our conversations on all things machine learning.

The acknowledgment section would not be complete without mentioning the many friends that I spent days and nights with, during my time at Purdue University. Thank you for supporting me when I needed it the most. To Saagar, thank you for exploring the food scene of West Lafayette with me. Nicoletta, Divya, and Arpan, I will never forget our cooking and movie nights, you made our office environment so much better. To my Lebanese friends, Bilal, Line, Tracy, and Elie for the endless laughs. To my fellow graduate students and roommates Rufat, Nyansafo, and Kola. To my Greek friends, Ekavi, Giwrgos, Katerina, Konstantinos, and Mantw, you will always have a special place in my heart, even if we are thousands of miles away.

I would also like to thank Alonso, you changed my life, and I consider you part of my family. Lastly, I want to thank my dad, Petros, for his continuous support during all these years. Dad, you are my rock during the best and toughest times, thank you for everything you have done for me.

Aspects of this work were partially funded by the U.S. Department of Defense through the Systems Engineering Research Center (SERC) under Contract HQ00034-13-D-0004 RT #206. SERC is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology.

TABLE OF CONTENTS

LIST OF TABLES	7
LIST OF FIGURES	10
ABSTRACT	16
1. INTRODUCTION	18
1.1 Research Background and Motivation.....	18
1.2 Risk, Systems Engineering Failures, and Failure Causes: Overview	23
1.3 Dissertation Outline and Contributions	28
2. EXPERIMENT I: CROWD-BASED RISK ASSESSMENT IN STUDENT PROJECTS ...	30
2.1 Experimental Setup and Design.....	30
2.2 Student Crowd Signals.....	32
2.3 Instructor Questions	40
3. EVALUATION OF PROJECT-BASED LEARNING (PBL) AND RECOMMENDATIONS FOR IMPROVEMENT	43
3.1 Failure Cause Occurrence: Comparing with Industry	43
3.2 PBL Improvement Recommendations	47
4. CROWD-BASED FAILURE PREDICTION IN STUDENT PROJECTS.....	56
4.1 Prediction Model Training.....	56
4.2 Prediction Model Validation.....	63
4.3 Prediction Model Reduction and Selection.....	64
5. EXPERIMENT II: TARGETED FEEDBACK TO PREVENT FAILURES IN SYSTEMS ENGINEERING	77
5.1 Experimental Setup and Design.....	77
5.2 Feedback process	80
5.3 Calculation of Overall Probability of Failure for a Project Team	82
5.4 Feedback Statement Development and Rules.....	85
5.5 Feedback Effectiveness Evaluation	94
6. COMPARISON OF CROWD SIGNALS BETWEEN EXPERIMENTS I AND II.....	101
7. CONCLUSIONS AND FUTURE WORK	137
7.1 Limitations	140

7.2	Suggestions for Future Research	141
7.2.1	Extension to student projects in multiple disciplines	141
7.2.2	Automated feedback process	141
7.2.3	Improving predictive models	142
7.2.4	Development of an integrated app environment.....	142
7.2.5	Industry setting	142
7.2.6	Introducing quantitative metrics of failure	143
7.2.7	Causal mechanisms between crowd signals and failure measures	143
APPENDIX A. CS–FC (“CROWD SIGNAL—FAILURE CAUSE”) CORRELATIONS		145
APPENDIX B. CONTINGENCY TABLES of FAILURE CAUSES.....		153
REFERENCES		156

LIST OF TABLES

Table 1: Definitions of project risk found in literature.	24
Table 2: Common causes of systems engineering failures that are observable in student projects. Adapted from (Sorenson and Marais, 2016) and (Aloisio, 2019).....	27
Table 3: Summary of data collection during academic year 2018–19. Student teams typically included 4-6 team members. The projects included both hardware and software deliverables as well as progress and final reports.....	32
Table 4: The questions that collected the crowd signals from the students. Each question was based on the definitions of corresponding literature.	34
Table 5: The questions to the instructors. Three questions captured occurrences of project failures and ten questions captured occurrences of failure causes.....	41
Table 6: Generic contingency table for each of the ten failure causes i.	45
Table 7: Occurrence measures and Barnard’s statistical test results for the failure causes. 1 out of 10 failure causes are underrepresented in the 28 student projects I studied compared to the 32 industry projects from [Sorenson and Marais, 2016 and Aloisio, 2019].	47
Table 8: The student questions that are actionable from the instructor with associated justification.	48
Table 9: Coding schemes for the 49 crowd signals , based on data type.....	52
Table 10: Mixed-effects logistic regression model coefficients for FC1: Failure to consider a design aspect.	54
Table 11: Instructor recommendations to improve student preparation in dealing with failure cause FC1: “Missing a design aspect” in PBL.	55
Table 12: Predictors and dependent variables for student project failure prediction. I built three models (one for each failure: budget, schedule, and technical requirements), from 51 predictors.	57
Table 13: Coding schemes for the input instructor measures, based on data type.	58
Table 14: Mixed-effects logistic regression model for prediction of budget failure.	60
Table 15: Mixed-effects logistic regression model for prediction of schedule failure.	61
Table 16: Mixed-effects logistic regression model for prediction of technical requirements failure.	62
Table 17: Generic confusion matrix for logistic regression models.	64
Table 18: Hybrid approach for model reduction and selection.....	66

Table 19: Budget model stepwise predictor variable removal. The process reduced the initial AIC by 68.....	68
Table 20: Budget model best subsets results. The table shows the best 15 models by AIC, from best to worst. The best model includes 10 predictor variables with the final model having an AIC of 264.75.	69
Table 21: Final budget model correlation coefficients.	70
Table 22: Schedule model stepwise predictor variable removal. The process reduced the initial AIC by 51.7.....	71
Table 23: Schedule model best subsets results. The table shows the best 15 models by AIC, from best to worst. The best model includes 15 predictor variables with the final model having an AIC of 315.5.	72
Table 24: Final schedule model correlation coefficients.	73
Table 25: Technical requirements model stepwise predictor variable removal. The process reduced the initial AIC by 61.26.	74
Table 26: Technical requirements model best subsets results. The table shows the best 15 models by AIC, from best to worst. The best model includes 12 predictor variables with a final model with an AIC of 271.12.....	75
Table 27: Final technical requirements model correlation coefficients.....	76
Table 28: Summary of data collection for experiment II during the Spring '21 semester. Student teams typically included 4-6 team members. The projects included both hardware and software deliverables as well as progress and final reports.	80
Table 29: The 35 feedback statements and associated rules for each existing correlation in the crowd signal–failure cause correlation matrix.	89
Table 30: Summary of feedback statements provided to each of the project teams during experiment II, for all weeks. Each team received three feedback statements from a pool of recommendations dependent on the treatment group they were a part of. The statements were repeated if the team did not provide new answers for a given week.	93
Table 31: Instructor evaluations at the end of semester during Experiment II. Instructors provided failure metrics for the 14 student projects. The project numbers do not correspond to the actual project team names for confidentiality purposes. One course was not viable for budget evaluation. “1” corresponds to failure and “0” to success.....	95
Table 32: Estimated sample failure proportions for budget, schedule, and requirements based on the instructor evaluation at the end of Experiment II.	96
Table 33: Barnard’s statistical test results for the targeted feedback. The statistical test suggests that the targeted feedback statements do not reduce the occurrence of failures in student projects, compared to the non-targeted feedback statements.	97
Table 34: The three additional questions that were part of the student survey during Experiment II to gauge how they received the feedback.	98

Table 35: Mixed-effects logistic regression model coefficients for FC2: Used inadequate justification.	146
Table 36: Mixed-effects logistic regression model coefficients for FC4: Lacked Experience... ..	147
Table 37: Mixed-effects logistic regression model coefficients for FC5: Kept poor records. ...	148
Table 38: Mixed-effects logistic regression model coefficients for FC6: Inadequately communicated.	149
Table 39: Mixed-effects logistic regression model coefficients for FC7: Subjected to inadequate testing.	150
Table 40: Mixed-effects logistic regression model coefficients for FC8: Managed risk poorly.	151
Table 41: Mixed-effects logistic regression model coefficients for FC10: Did not allow system aspect to stabilize.	152
Table 42: Contingency table for FC1: Failed to consider a design aspect.	153
Table 43: Contingency table for FC2: Used inadequate justification.	153
Table 44: Contingency table for FC3: Failed to form a contingency plan.	153
Table 45: Contingency table for FC4: Lacked experience.	153
Table 46: Contingency table for FC5: Kept poor records.	154
Table 47: Contingency table for FC6: Inadequately communicated.	154
Table 48: Contingency table for FC7: Subjected to inadequate testing.	154
Table 49: Contingency table for FC8: Managed risk poorly.	154
Table 50: Contingency table for FC9: Violated procedures.	154
Table 51: Contingency table for FC10: Did not allow system aspect to stabilize.	155

LIST OF FIGURES

Figure 1: Historical project success rates for 2011–2020, adapted from Project Management Institute (PMI) Pulse of the Profession yearly reports. The most prominent problem for projects is completing their milestones on time, with almost half of them failing in terms of this metric. The success rates for all metrics have not improved much since 2011, perhaps indicating that we may have reached a plateau.	20
Figure 2: Failures and failure causes are represented as a causal relationship (solid line). Crowd signals correlate with the occurrence of failure causes and failures (dashed line). I use the correlation relationship (a) to build predictive models that alert of future failures in student teams and correlation relationship (b) to develop targeted feedback to reduce failures by addressing the underlying failure causes.	21
Figure 3: The Crowd-based Risk Assessment Prototype that uses the crowd signals as inputs to predict future failures and then suggests targeted feedback, based on correlations between the failure causes and crowd signals.	23
Figure 4: Dissertation outline and contents.	29
Figure 5: Recruitment flyer as distributed to the students during the recruitment process of the first experiment. The number of gifts cards (X) and value of each (Y) changed per semester: initially (Summer 2018) I included two gift cards of \$50 that were distributed at the end of the semester once. In subsequent semesters (Fall 2018/Spring 2019) I switched to a weekly \$20 gift card model because I wanted to encourage participation every week.	31
Figure 6: 10-fold cross validation process. The 267 data points (initial observations were 304, including 37 NAs) were split in 10 folds of 27 or 26. At each iteration, 240 or 241 data points were used as the training set for the logistic regression models and then the remaining 27 or 26 data points as the testing set. I recorded how many correct predictions (of the 27 or 26) the algorithm correctly identified in each iteration. I repeated the process until all observations had the chance to be included in the testing fold.	63
Figure 7: All three prediction models correctly predicted, on average, 60 to 65% of outcomes of unknown data. The schedule model had the highest variance of the three between the folds.	64
Figure 8: Relative question importance based on their inclusion in the three final reduced models. Light grey-coded inputs appeared in 1/3 final models and dark-grey coded inputs appeared in 2/3 final models. Previous project status was the only input variable that appeared in all three final predictive models.	68
Figure 9: The final best budget model is, on average, more accurate ($73.11 \pm 6.92\%$) and predicts correctly with less variance than the initial budget model ($64.50 \pm 9.96\%$).	70
Figure 10: The final best schedule model is, on average, more accurate ($75.27\% \pm 9.21\%$) and predicts correctly with less variance than the initial model ($60.38\% \pm 13.64\%$).	73

Figure 11: The final best technical requirements model is, on average, more accurate ($76.71 \pm 6.90\%$) and predicts correctly with less variance than the initial model ($66.31 \pm 10.32\%$).	76
Figure 12: Recruitment flyer as distributed to the students during the recruitment process of Experiment II. 10 random students won a \$20 gift card.	78
Figure 13: The two treatment groups used in Experiment II. The difference was the process for the feedback statements: one group received from statements that applied to them based on the rules, whereas the other group received from statements that did not apply to them based on the rules.	79
Figure 14: The dynamic feedback process. At each week t , through the failure predictive models, I calculated the probability of failure for each metric and team for the following week $t+1$. Using this probability and the associated feedback rules, I provided recommendations to the team members.	82
Figure 15: The crowd signal–failure cause correlation matrix. “+” indicates a positive correlation, “-” indicates a negative correlation, and “0” indicates no correlation. When the questions have categorical answers, the correlation is labeled with the answer that it corresponds to. FC3 and FC9 are excluded due to low occurrence ratios in the corresponding model training data sets.	86
Figure 16: 68% of the responses (34 out of 50) from teams that received targeted feedback show that the students changed their behavior, compared to 34.8% of the responses (16 out of 46) from teams that received the non-targeted feedback.	99
Figure 17: Teams that received non-targeted feedback rated, most of the time, the feedback statements as moderately helpful, with more responses (15) scoring it as 1 or 2 than 4 or 5 (13). Responses from teams that received targeted feedback are towards higher ratings, with most responses (15 out of 50) rating the statements as very helpful.	99
Figure 18: Both treatment groups thought the feedback would have some positive impact on their project, possibly due to the positive nature of the statements. The responses from the targeted feedback teams had a larger representation in the >75% chance of positive impact answer and smaller representation in the <25% positive impact answer, compared to the non-targeted treatment group. The “TF” group had 4 more answers than the “NTF” group, which contributes to the observation.	100
Figure 19: Experience level of the $N = 74$ students that have participated in experiment I. I measured experience based on the number of engineering projects the students have been part of in the past. Most students had previous experience from 1–4 projects.	101
Figure 20: Most of the $N = 74$ students that participated in Experiment I, got involved, on average, at most one time with tasks outside their immediate responsibilities.	102
Figure 21: Most of the $N = 53$ students that participated in Experiment II, got involved, on average, between 1 to 2 times with tasks outside their immediate responsibilities.	102
Figure 22: Statistics of the student responses to Q3. The majority of responses showed that students were sometimes unable to focus on their projects.	103

Figure 23: Statistics of the student responses to Q4. For both experiments most showed that there was some level of coordination even when working on separate tasks, but students in Experiment II interacted more often with each other.	104
Figure 24: Statistics of the student responses to Q5. For both experiments most responses showed there was some level of meaningful progress during a typical week, but students in Experiment II made such progress more frequently.	104
Figure 25: Of the N = 74 students who participated in Experiment I, most knew at least what half of their team members were working on.	105
Figure 26: During experiment II, students did not know as much about their team members' activities as during Experiment I. Hybrid learning and safety measures due to the COVID-19 pandemic may be responsible for that, given students have to follow capacity limits in their workspaces so they could not work in the same room as much.	105
Figure 27: Statistics of the responses to Q7. Only two responses during Experiment II indicated students had no freedom at all when completing a project task. More than 75% of responses showed that the instructors for the courses in both experiments gave students at least moderate freedom on how to complete their project objectives.	106
Figure 28: During Experiment I, students gave varying chances of success if their teams would have to complete the projects without supervision for the remaining of the semester. The responses to this question likely depend on the project phase, the student's confidence in their own and their team's capabilities, and overall knowledge about the project's next steps.	107
Figure 29: During Experiment II, the majority of the responses from students appear to get grouped around two options: 1) students that think they need the instructor supervision and would only have 30–60% of success without it, and 2) students that think they could be successful (>70%) without the instructor.	107
Figure 30: The percentage of tasks that can be performed independently of the rest of the project resembles a uniform profile for Experiment I. The degree of modularity in a project is likely dependent on project phase and current requirements that continuously change and get updated throughout the semester.	108
Figure 31: Similar to Experiment I, the percentage of tasks that can be performed independently of the rest of the project resembles a uniform profile for Experiment II.	108
Figure 32: Statistics of the responses to Q11. More than 75% responses indicate that the project objectives were at least moderately clear for both experiments. The objectives were likely set by the instructor, who directly asked for a task to be completed and gave specific milestones for the students. Also, some teams may have set their own objectives, since they may have some freedom on how to complete the project tasks.	109
Figure 33: Statistics of the responses to Q12. More than 80% of responses indicate that students would at least consider working on the project with an entirely new team for both experiments. 12% of responses for Experiment II show that students would definitely abandon the project (compared to only 5% in Experiment I). The results are likely related to the quality of the courses, the value the students saw in participating, and how long they were part of the project.	110

Figure 34: Statistics of the responses to Q13. More than 85% of responses indicate at least moderate availability of resources to use in a given week. The quality of the laboratories and campus resources are likely related to the response profile for this question. Very few responses communicated very low availability, perhaps due to a specific need or request for the project that was not met.	111
Figure 35: Statistics of the responses to Q14 (averaged by all N = 304 responses) during Experiment I. 41.1% of responses indicate that sometimes there is lack of communication while students work together. 6.9% said that they always noticed a silent room, which may be a sign of poor team cohesion. Poor communication is a frequent problem amongst teams in engineering projects, and student projects also confirm that.	112
Figure 36: Statistics of the responses to Q15. The majority of students rarely get frustrated with each other and their team.	113
Figure 37: Statistics of the responses to Q16. For Experiment I, the majority of responses reflected that students “sometimes” come up with or agree to new ideas for the projects. For Experiment II, the responses showed a more balanced profile compared to Experiment I.	114
Figure 38: Statistics of the responses to Q17. The majority of students would rarely or never skip or cancel an obligation or task in a given week.	115
Figure 39: Statistics of the responses to Q18. A relatively small number of responses showed a student to always be the center of attention.	116
Figure 40: Statistics of the responses to Q19. 19.1% of responses in Experiment I and 13% in Experiment II that students never had one of their team members share important things about their life with them, while only 6.2% for Experiment I and 2% for Experiment II said they always did.	117
Figure 41: Student confidence in their spending estimate for Experiment I. Most responses concentrated around 40–50% and 70–100%.	118
Figure 42: Student confidence in their spending estimate for Experiment II. In 38 instances, students gave a spending estimate with absolute confidence.	118
Figure 43: Statistics of the responses to Q26. The majority of students said their teams handled problems appropriately.	119
Figure 44: Statistics of the responses to Q27. Students during Experiment II did better risk management than students during Experiment I, by considering new risks to project updates. The feedback statements during Experiment II, a lot of which guide students to avoid the failure causes, may have contributed to this result.	120
Figure 45: Statistics of the responses to Q28. At least 35% of responses indicated that students were limited by processes or rules outside their control.	121
Figure 46: Statistics of the responses to Q29. During Experiment I, students were frustrated more frequently due to bureaucracy or rules.	121
Figure 47: Statistics of the responses to Q30. Most teams likely went through all three options (decrease, no change, increase) related to the number of outputs they produced, depending on	

project phase. The responses show some balanced split between some increase (23% for Experiment I, 25% for Experiment II) and some decrease (14% for Experiment I and 18% for Experiment II).....	122
Figure 48: Statistics of the responses to Q31 during Experiment I. 85% of responses came from students who spent between 1 to 3 hours on social media, while the remaining 15% spent more than 3 hours.....	123
Figure 49: Statistics of the responses to Q32 during Experiment I. Based on the responses, students chose to eat fast food slightly more frequently than dining halls. The majority opted for home-prepared meals.	124
Figure 50: Statistics of the responses to Q33 during Experiment I. 72% of responses indicate that students had breakfast, at least a few times in a given week. The courses we collected data from were held in the morning, and occur 2–3 times per week.	124
Figure 51: Most students thought about their projects between 20–50% of their working time during Experiment I, on average.....	125
Figure 52: Similarly to Experiment I, most students thought about their projects between 20–50% of their working time during Experiment II, on average.	125
Figure 53: Most students, on average, met up to 2 times with their teams to work on their project outside regular class times during Experiment I, which is often necessary to complete demanding tasks.....	126
Figure 54: For Experiment II, there was larger number of teams that met outside class time for 2 or more times in a week. It is possible that students were making up for the lack of on-campus access due to COVID-19 restrictions.....	126
Figure 55: Most students ordered, on average, between 0–5 new parts during Experiment I, with some occasions where a larger order was necessary. These larger orders were likely sets of equipment like screws or bolts.....	127
Figure 56: For Experiment II, the number of ordered parts per week is similar to Experiment I. Most teams ordinarily do not need to order new tools frequently.	127
Figure 57: Majority of responses (75%) indicate that students exercised at least one time in a given week. Frequent physical exercise may correlate with improved individual performance and intellectual ability, that may influence how well students do on technical projects.....	128
Figure 58: When ranking possible mishaps from highest to lowest risk, students considered missing their technical requirements as the highest risk during Experiment I and Experiment II, followed by a schedule mishap, and lastly a cost mishap. Budget likely mattered the least for them because they did not monitor it closely and likely did not know what the overall budget was. In contrary, they knew the timeline and requirements they had to meet, and they likely suspected they are also evaluated based on these two metrics more than how well they follow a budget.	129
Figure 59: When asked to pick a failure they would prefer associated with the three metrics, the students gave more weight on avoiding a technical requirements failure, and most said they would rather have a budget failure.....	130

Figure 60: Statistics of the responses to Q41. 35% of responses during Experiment I indicated disagreement to new ideas when lacking understanding about potential implications, compared to 23% during Experiment II.	131
Figure 61: Statistics of the responses to Q42. Responses were similar during both experiments, majority of responses showed students not arguing during 2/3 typical weeks.	132
Figure 62: Statistics of the responses to Q43. Responses were similar during both experiments, with an almost perfect split between students who did and did not identify a single project decision as the most important.	132
Figure 63: Statistics of the responses to Q44. 7% (Experiment I) and 68% (Experiment II) of responses show students spent time thinking about what might go wrong in their project, indicating PBL helped students develop risk management skills.	133
Figure 64: Statistics of the responses to Q45. 45% (Experiment I) and 68% (Experiment II) talked with other colleagues and got ideas from other teams.	134
Figure 65: Statistics of the responses to Q46. Almost for half their time involved in the projects, students learned something new.	134
Figure 66: Statistics of the responses to Q47. The results show mostly equal time efficiency between the two experiments.	135
Figure 67: More than 2/3 of the responses come from students who said they thought through all solutions before making a decision during both experiments, with some indicating the question did not apply for a particular week, suggesting that there was no project decision for them to make.	136
Figure 68: Industry prototype for project failure prediction and prevention.	143

ABSTRACT

Most engineering curricula in the United States include some form of major design project experiences for students, such as capstone courses or design-build-fly projects. Such courses are examples of project-based learning (PBL). Part of PBL is to prepare students—and future engineers—to deal with and prevent common project failures such as missing requirements, overspending, and schedule delays. *But how well are students performing once they join the workforce?* Unfortunately, despite our best efforts to prepare future engineers as best we can, the frequency of failures of complex projects shows no signs of decreasing. In 2020 only 53% of projects were on time, 59% within budget, and 69% met their goal, as reported by the Project Management Institute. If we want to improve success rates in industry projects, letting students get the most out of their PBL experience and be better prepared to deal with project failures before they join the workforce may be a viable starting point.

The overarching goal of this dissertation is to identify and suggest improvements to areas that PBL lacks when it comes to preparing students for failure, to investigate student behaviors that lead to project failures, and to improve these behaviors by providing helpful feedback to students.

To investigate the actions and behaviors that lead to events that cause failures in student projects, I introduced “crowd signals”, which are data collected directly from the students that are part of a project team. In total, I developed 49 survey questions that collect these crowd signals. To complete the first part of the dissertation, I conducted a first experiment with 28 student teams and their instructors in two aerospace engineering PBL courses at Purdue University. The student teams were working on aircraft designs or low-gravity experiments.

Does PBL provide sufficient opportunities for students to fail safely, and learn from the experience? How can we improve? To identify areas that PBL may lack, I compared industry failure cause occurrence rates with similar rates from student teams in PBL courses, and then provided recommendations to PBL instructors. Failure causes refer to events that frequently preceded budget, schedule, or requirements failures in industry, and are identified from the literature. Through this analysis, I found that PBL does not prepare students sufficiently for situations where

the failure cause *missing a design aspect* occurs. The failure cause is fundamentally linked to proper systems engineering: it represents a scenario where, for example, students failed to consider an important requirement during system development, or did not detect a design flaw, or component incompatibility. I provided four recommendations to instructors who want to give their students more opportunities to learn from this failure cause, so they are better prepared to tackle it as engineers.

Is crowdsourced information from project team members a good indicator of future failure occurrences in student projects? I developed models that predict the occurrence of future budget, schedule, or requirements failures, using crowd signals and other information as inputs, and interpreted those models to get an insight on which student actions are likely to lead to project failures. The final models correctly predict, on average, $73.11 \pm 6.92\%$ of budget outcomes, $75.27\% \pm 9.21\%$ of schedule outcomes, and $76.71 \pm 6.90\%$ of technical requirements outcomes. The previous status of the project is the only input variable that appeared to be important in all three final predictive models for all three metrics. Overall, crowdsourced information is a useful source of knowledge to assess likelihood of future failures in student projects.

Does targeted feedback that addresses the failure causes help reduce failures in student projects? To improve student behaviors that lead to project failures, I used correlations between failure measures and the crowd signals as a guide to generate 35 feedback statements. To evaluate whether the feedback statements help reduce project failures in the student teams, I conducted a second experiment at Purdue University with 14 student teams and their instructors. The student teams were enrolled in aircraft design, satellite design, or propulsion DBT courses. The student teams were split in two treatment groups: teams that received targeted feedback (i.e., feedback that aimed to address the failure causes that the specific team is most prone to) and teams that received non-targeted feedback (i.e., feedback that is positive, but does not necessarily address the failure causes the specific team is most prone to). Through my analysis, I found that my targeted feedback does not reduce the failure occurrences in terms of any metrics, compared to the non-targeted feedback. However, qualitative evaluations from the students indicated that student teams who received targeted feedback made more changes to their behaviors and thought the feedback was more helpful, compared to the student teams who received non-targeted feedback.

1. INTRODUCTION

1.1 Research Background and Motivation

Most engineering curricula in the United States include some form of major design project experiences for students, such as capstone courses or design-build-fly projects. Such courses are examples of project-based learning (PBL). Project-based learning is the theory and practice of using real-world projects that have time restrictions, engineering constraints, specific objectives, and aim to facilitate individual and collective learning [DeFillippi, 2001]. PBL is a learner-centered approach that allows students to engage with an ill-defined project to promote research, teamwork, critical thinking, and synthesis of multidisciplinary technical knowledge [Mills and Treagust, 2003; Savery, 2006]. The instructor usually acts as a facilitator who guides the students through the learning experience as necessary, while allowing them to take responsibility for their project decisions [Atman et al., 2007]. PBL is widely considered to be successful, with students positively evaluating the approach and suggesting that it helps them develop their engineering intuition, makes them responsible for their decisions, and to become flexible thinkers [Hall et al., 2012; Frank et al., 2003]. Researchers and faculty also consider PBL an integral part of education that teaches students to handle complex problems that require diverse thinking and integration skills [Lehmann et al., 2008].

Despite the abundance of literature about PBL (e.g., see Kokotsaki et al. (2016) for a review), research does not appear to have influenced PBL in practice [McCormick et al., 2013; Cottrell, 2006]. One reason for this disparity is that research studies focus on theoretical problems designed by education professionals to exactly meet PBL criteria, while faculty create design projects based on subject-matter expertise, likely without exposure to most recent PBL research on how to maximize students' conceptual understanding of engineering design [John and Thomas, 2000]. Researchers in the field also question whether we have been evaluating the effectiveness of PBL correctly, emphasizing that instead of evaluating PBL based on student exam scores, we need strategies to identify whether particular aspects of PBL help educators cater to specific student learning outcomes [Hmelo-Silver, 2004].

Some of these learning outcomes specifically refer to students being capable of applying engineering design knowledge to produce solutions that meet needs in regard to safety, requirements, and economic factors (ABET criterion 3.2, 2019–2020). Therefore, part of PBL is to also prepare students—and future engineers—to deal with and prevent common project failures such as missing requirements, overspending, and schedule delays. The first part of this dissertation evaluates whether PBL provides sufficient opportunities for students to fail safely, learn from the experience, and thereby be more prepared to avoid failure in professional practice.

If engineering training in general, and PBL in particular, was achieving its mission, we would hope to see a decrease in project failures. Unfortunately, despite our best efforts to prepare future engineers as best we can, the frequency of failures of complex engineering projects shows no signs of decreasing. Of the 72 major United States defense programs in progress in 2008, only eleven of them were on time, on budget, and met performance criteria [Charette, 2008]. The problems for U.S. aerospace and defense programs have only worsened since then: total cost overruns “have risen from 28 percent to 48 percent, from 2007 through 2015” [Lineberger and Hussein, 2016]. In a recent assessment of U.S. Defense Acquisitions, the U.S. Government Accountability Office (GAO) found that these programs were “not yet fully following a knowledge-based acquisition approach”, which will result in “cost growth or schedule delays” [GAO, 2017]. The consumer goods sector has also had many examples of product failures, such as the Xbox 360 “Red Rings of Death” [Takahashi, 2008] or the Ford Explorer rollover problems [Bradsher, 2000].

Figure 1 shows historical information on project success rates since 2011, confirming that project performance has not improved much in recent years. In 2020 only 53% of projects were on time, 59% within budget, and 69% met their goals [Project Management Institute, 2020, p.12].

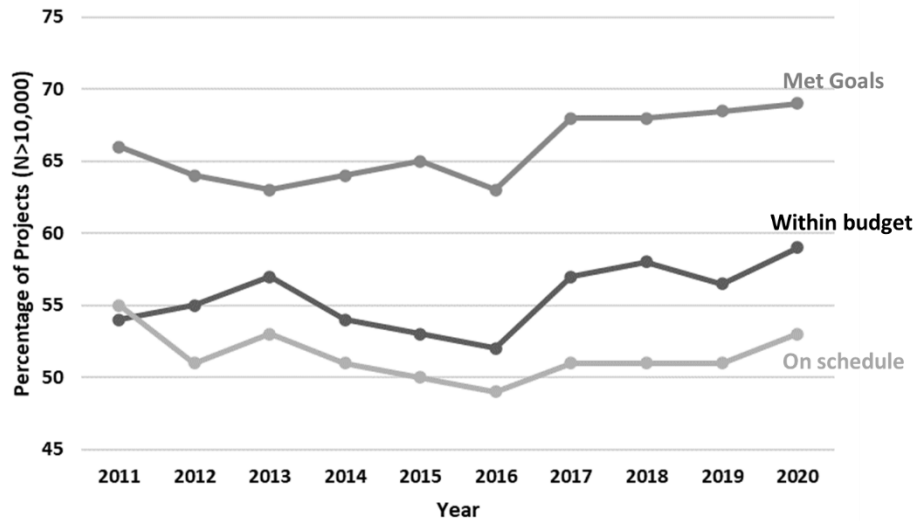


Figure 1: Historical project success rates for 2011–2020, adapted from Project Management Institute (PMI) Pulse of the Profession yearly reports. The most prominent problem for projects is completing their milestones on time, with almost half of them failing in terms of this metric. The success rates for all metrics have not improved much since 2011, perhaps indicating that we may have reached a plateau.

The high frequency of failures has long served as a motivation of systems engineering research to investigate the root causes of these phenomena. Part of the systems engineering process involves assessing risk of project failures like cost overruns, schedule slips, and failure to meet technical requirements [Bahill and Dean, 1996]. Sorenson and Marais (2016) looked for the underlying causes (referred to as “failure causes”) of 62 systems engineering failures and accidents, and found that even in new, one-of-a-kind high-tech systems, failures do not involve previously unknown phenomena or black swans, but rather prosaic and predictable white swans.

Given the occurrences of these failure causes in industry and the need to improve engineering training overall, PBL may be a viable starting point to study and improve the behavioral patterns of students before they join the workforce and default to actions that lead to project failures.

To further understand *why* these failure causes happen, I leverage the idea that the actions and behaviors of students in project teams may relate to the occurrences of failure causes, which eventually lead to project failures. Therefore, asking project members directly about what they are doing or thinking may be a valid approach to predict and prevent future failures. My approach

introduces “crowd signals” to capture the human actions and behaviors that lead to failure causes and eventually to failures. Crowd signals are data collected directly from the students that are part of a project team. By collecting risk information directly from team members, the goal is to capture project risk that comes from humans at its source, and to do so frequently, continuously, and in an efficient manner.

I use the crowd signals to (a) predict upcoming failures and (b) identify which behaviors, actions, biases, or other characteristics correlate with specific failure causes, which helps in developing feedback statements (Figure 2).

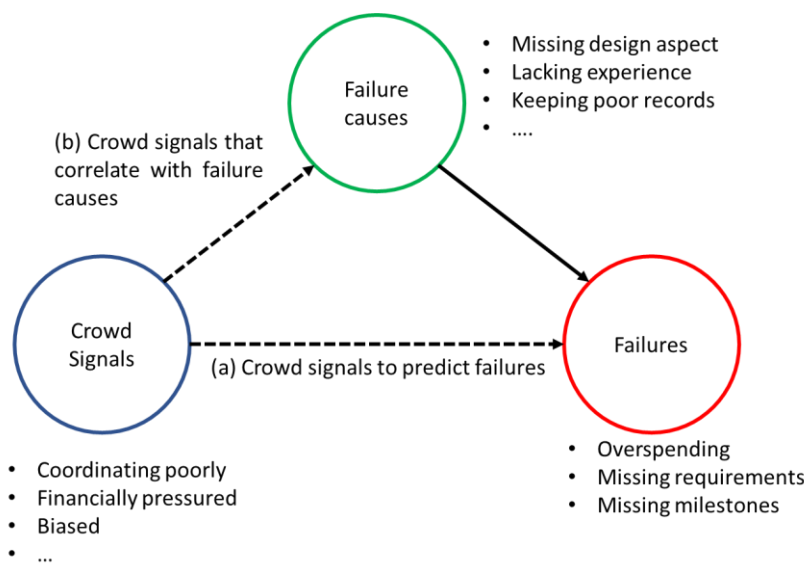


Figure 2: Failures and failure causes are represented as a causal relationship (solid line). Crowd signals correlate with the occurrence of failure causes and failures (dashed line). I use the correlation relationship (a) to build predictive models that alert of future failures in student teams and correlation relationship (b) to develop targeted feedback to reduce failures by addressing the underlying failure causes.

The crowd signals, together with observations of failure from the instructor, provide all the necessary information to train models that predict the probability of occurrence of a project failure. The crowd signals are also useful to find correlations between specific human actions or behaviors and the occurrence of failure causes, which motivates the development of targeted feedback to address these failure causes before they lead to failures. For example, if we knew that more frequent team meetings correlate with reduced occurrences of the failure cause *Inadequate*

Communication, then we could provide recommendations to a student team that does not meet often to help them improve their communication and avoid this failure cause.

Machine learning techniques can help process the crowd signals, build predictive models, and find correlations between the crowd signals, failures, and failure causes. Machine learning can combine inputs from multiple sources and uncover hidden patterns, not all of which may be predictable a priori, and some of which may differ between teams or settings. In this dissertation, I used logistic regression to investigate all the correlations and build all necessary models shown in Figure 2.

The main goal of the dissertation is to identify and suggest improvements to areas that PBL lacks when it comes to preparing students for failure, to evaluate whether crowdsourced information from the project team members can help to predict future failures, and to test whether targeted feedback can prevent these failures in student projects.

The dissertation considers three main research questions:

1. *Does PBL provide sufficient opportunities for students to fail safely, and learn from the experience? How can we improve?*

To answer this question, I compared failure cause occurrence rates from industry project failures with similar rates from student teams in PBL courses. Then, I used the correlations between the crowd signals and the underrepresented failure cause to provide four recommendations to instructors.

2. *Is crowdsourced information from project team members a good indicator of future failure occurrences in student projects?*

To answer this second question, I developed models that predict the occurrence of future failures, using crowd signals and other information as inputs, and evaluated those models.

3. *Does targeted feedback that addresses the failure causes help reduce failures in student projects?*

To answer the last question, I used the correlations between all of the failure causes and crowd signals as a guide to generate targeted feedback that aimed to address the failure causes that the

specific team is most prone to. Then, I compared the occurrence of failures between student teams that received the targeted feedback and student teams that received non-targeted feedback. Also, I used qualitative evaluations from the students to gauge their opinions between the targeted and non-targeted feedback.

To answer the research questions, I conducted two separate experiments in PBL courses that include complex engineering projects (i.e., senior design, capstone, or design-build-fly) at Purdue University. The experiments involved surveying both instructors and students. The students were the project team, and provided the crowd signal inputs during the time they worked on their design projects. The instructors played the role of management for the student teams because they were the primary stakeholders and (should) closely monitor the student projects.

Figure 3 shows an overview of the crowd-based failure prediction and prevention prototype that I developed as part of answering research questions (ii) and (iii) of this dissertation. The heart of the prototype is a family of logistic regression machine learning algorithms (*Predictive Logistic Regression Algorithms*) that process the collected signals from the project teams (*Crowd Signals*) and make predictions about future failures (function a). The crowd signals are collected using mobile- and web-accessible surveys (*Qualtrics survey*). With knowledge of future risk, the prototype provides feedback to address the project and team actions before they lead to failure (function b).

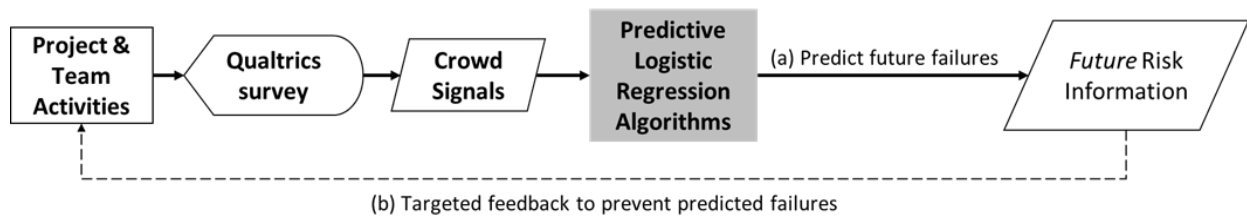


Figure 3: The Crowd-based Risk Assessment Prototype that uses the crowd signals as inputs to predict future failures and then suggests targeted feedback, based on correlations between the failure causes and crowd signals.

1.2 Risk, Systems Engineering Failures, and Failure Causes: Overview

The dissertation builds on literature and terminology from project risk management and systems engineering. To provide context for the reader, this section summarizes notable literature on these topics.

The word “risk” is a prime example of a word humans frequently use in our daily conversations, but that each of us understands, interprets, and reacts to differently [Sjöberg, 2000]. There are multiple different uses of the word in the literature [Al-Bahar and Crandall, 1990]. My focus for this research is on risk associated with project performance (“project risk”), primarily regarding the likelihood of systems engineering failures. Table 1 shows seven definitions of project risk from literature. Definitions 1–5 come from project management handbooks or standards, while definitions 6–8 come from research articles.

Table 1: Definitions of project risk found in literature.

#	Definition of project risk	Sources
1	“Effect of uncertainty on objectives”	International Organization for Standardization ISO 31000:2018, 2018, Risk Management—Principles and Guidelines (section 3.1)
2	“An uncertain event or condition that, if it occurs, has a positive or negative affect on a project’s objectives.”	Hillson, 2014
3	“Events with a negative impact represent risks, which can prevent value creation or erode existing value.”	Committee of Sponsoring Organizations of the Treadway Commission (COSO), 2004, (p. 1)
4	“A possible occurrence which could affect (positively or negatively) the achievement of the objectives for the investment.	Institution of Civil Engineers, Risk Analysis and Management for Projects (RAMP), 2014, (Appendix 1, p. 74)
5	“The chance of something happening that will have an impact upon objectives. It is measured in terms of consequences and likelihood.”	Joint Australian/New Zealand Standard AS/NZS 4360:1999, 1999, (p.3)
6	“Undesired events that may cause delays, excessive spending, unsatisfactory project results, safety or environmental hazards, and even total failure [...]”	Raz et al. 2002
7	“A set of factors or conditions that can pose a serious threat to the successful completion of a software project.”	Wallace et al. 2004
8	A project activity with high likelihood of adverse result, small ability to influence it, and severe consequences.	Keizera et al. 2002

“Systemic risk” originated in the financial sector to describe catastrophic events that could collapse an entire sector or market. Kaufman and Scott (2003) generalized the term in the context of a complex project as “a risk that originates from multiple sources, affects multiple agents and propagates quickly among individual parts or components of the network.” Systemic risk goes beyond the more traditional narrow view of risk and includes additional aspects. Perhaps this definition is more suitable in the context of today’s large and complex technical projects (e.g., construction or aerospace engineering projects) that include risk from a variety of sources that correlate with unknown mechanisms [Gandhi and Gorod, 2012; Kremljak and Kafol, 2014].

There are three main approaches that researchers and professionals follow to assess risk: qualitative, quantitative, and semi-quantitative. Perhaps the most widely known quantitative method is to evaluate risk as the product of the likelihood and impact, also known as expected value theory [Al-Bahar and Crandall, 1990; Williams, 1996]. Others have used different formulas as part of a quantitative calculation, such as including a discrimination factor to represent the impact of the risk to the overall project [Cervone, 2006], including a detection factor to represent the likelihood of correctly identifying risk [Carbone and Tippet, 2004], or other deterministic approaches [Muriana and Vizzini, 2017]. Qualitative methods for risk communication include Failures Modes and Effects Analysis (FMEA) [Bouti and Kadi, 1994], risk matrices that describe the likelihood and impact in qualitative terms, and fuzzy logic [Carr and Tah, 2001]. Some authors have pointed out that sometimes qualitative methods can be vague because they use definitions such as “high” or “low” that may not mean the same thing when evaluated by different parties [Tah and Carr, 2001]. Lastly, semi-quantitative methods include both quantitative and qualitative elements. For example, qualitative matrices of risk dimensions followed by some form of numerical calculations for ranking and assessing risk [Cooper et al., 2005; Yoon et al., 2014], or Bayesian Belief Networks (BBNs) [Fan and Yu, 2004; Lee et al., 2009], or computational decision frameworks [Fang and Marle, 2012].

When discussing project performance, it is common practice to form a set of *project success criteria* that determine how to evaluate a project. There is no universal agreement on what the project success criteria should be, in fact, success criteria have changed over the years [Lim and Mohamed, 1999; Ika, 2009]. Projects can differ in so many ways from one another (industry,

stakeholders, size, scope, budget, schedule, location, etc.) that it is near impossible to come up with a unique set of success criteria that applies to all projects. Also, deciding which group of stakeholders selects the success criteria is not straightforward: the CEO, customers, the government, and external contractors are just some of the viable candidates. For example, the government may consider a project successful if it adheres to all regulations and policy, while customers may give more weight to the quality and functionality of the product. Lastly, project success may change during different project phases. For example, the design phase of a large aerospace vehicle might be evaluated based on technical objectives, while the manufacturing phase might be evaluated mainly on cost, schedule, and quality. In the literature, the most common basis of project success is the “iron triangle”, which considers project cost, schedule, and quality [Chua et al., 1999]. Others have incorporated client satisfaction [White and Fortune, 2002; Aloini et al., 2007; Zwikaël and Ahn, 2011], or technical and project performance [Shenhar et al., 2002] in their criteria. For this work, like PMI, I consider three criteria of project evaluation: cost, schedule, and technical requirements.

Current risk assessment approaches have not helped as much to reduce the project success rates in recent years, as shown earlier (Figure 1). It is difficult to identify and properly assess risk in a timely manner for a wide range of reasons, including system complexity [Abt et al., 2010], lack of management insight [Yoon et al., 2014], and lack of past evidence [Apostolakis, 2004]. Also, existing risk assessment methods often rely on intuition, hindsight, and experience of failures, which may not always be trustworthy or available [Frosdick, 1997].

When investigators study past failures in complex systems, they often encounter multiple and interrelated causes that may not necessarily cause a failure on their own [Paté-Cornell, 1993; Rasmussen 1997]. To demonstrate the variety and interrelation of failure causes in complex projects, consider the Boeing 787. The project was three years behind schedule and over budget before delivery to the launch customer. Investigation revealed issues with improper testing of the batteries, ineffective outsourcing, supply chain communication issues, and poor management decisions [Denning (for Forbes), 2013]. Previous research has identified some of the more frequent causes that lead to these types of failures [Sorenson and Marais, 2016]. In this work, I consider 10 failure causes that are observable in student projects (shown in Table 2).

*Table 2: Common causes of systems engineering failures that are observable in student projects.
Adapted from (Sorenson and Marais, 2016) and (Aloisio, 2019).*

<i>Systems Engineering Failure Causes</i>		<i>Explanation</i>
FC1	Failed to consider design aspect	Students failed to consider an aspect in the system design. In many cases, this causal action describes a design flaw, such as a single-point failure or component compatibility.
FC2	Used inadequate justification	Students used inadequate justification for a decision.
FC3	Failed to form a contingency plan	Students failed to form a contingency plan to implement if an unplanned event occurred.
FC4	Lacked experience	Students' lack of experience or knowledge led to the failure.
FC5	Kept poor records	Students did not review documentation or other work sufficiently to capture errors and deficiencies.
FC6	Inadequately communicated	Students failed to communicate with each other such that they were confused with the information they were given, had to "fill in the gaps" in the information they were given, or were not notified about important information at all.
FC7	Subjected to inadequate testing	One or more students subjected a component or subsystem to inadequate testing. This causal action captures inadequate tests as well as adequate tests performed inadequately.
FC8	Managed risk poorly	Students failed to identify, assess, formulate, or implement a proper mitigation measure.
FC9	Violated procedures	Students violated a procedure pertaining to the system, such as a maintenance or operation procedure.
FC10	Did not allow system aspect to stabilize	Students did not allow a system aspect like design or requirements to stabilize before moving forward with an action.

I consider these failure causes to contribute to the overall risk of a student project by increasing the likelihood of a project failure occurring. For example, a project may remain "on schedule" for months, until the occurrence of several of these failure causes delay the project. In this manner, the reader can think of these failure causes as "increasing risk events": events that increase the overall failure risk of a project. Apart from the causal relationship between failures and failure causes, researchers also recognize that human behavior and characteristics are in some way responsible for the occurrences of most failure causes and failures, but we may not know exactly how, depending on the situation. People (usually) do not willfully conduct inadequate testing or

omit important design aspects but can be careless or lose focus while performing a critical design or testing task. Research has identified activities, behaviors, and even personality characteristics that can lead to poor habits and performance. For example, Halfhill et al. (2005) found that military teams with high levels of both conscientiousness and agreeableness received higher performance ratings than other teams.

1.3 Dissertation Outline and Contributions

The remaining of this dissertation is organized as follows. Chapter 2 is an overview of the first experiment and includes the experimental design and recruiting process, the student crowd signals, and the instructor questions. Chapter 3 is an evaluation of project-based learning based on the data from the first experiment and failure cause data from industry. Chapter 3 also includes recommendations to instructors to better prepare students based on areas that I identified PBL is lacking. Chapter 4 provides information on the logistic regression failure prediction models and rationale, and discusses model validation and model selection. Chapter 5 focuses on the second experiment and describes the procedures, the targeted feedback process, and the findings on whether the targeted feedback is helpful in reducing failures in student projects. Chapter 6 shows descriptive statistics of the student responses to the questionnaire that collected the crowd signals, showing some comparisons between the two experiments. The document concludes with Chapter 7, which is a summary of the research conclusions, the limitations, and some ideas for future work.

Chapter 1	Introduction	<ol style="list-style-type: none"> 1. Research Background and Motivation 2. Risk, Systems Engineering Failures, and Failure Causes: Overview 3. Dissertation Outline
Chapter 2	Experiment I: Crowd-Based Risk Assessment in Student Projects	<ol style="list-style-type: none"> 1. Experimental Setup and Design 2. Student Crowd Signals 3. Instructor Questions
Chapter 3	Evaluation of Project-Based Learning (PBL)	<ol style="list-style-type: none"> 1. Failure Cause Occurrence: Comparing with Industry 2. PBL Improvement Recommendations
Chapter 4	Crowd-based Failure Prediction in Student Projects	<ol style="list-style-type: none"> 1. Prediction Model Training 2. Prediction Model Validation 3. Prediction Model Reduction
Chapter 5	Experiment II: Targeted Feedback to Prevent Failures in Student Projects	<ol style="list-style-type: none"> 1. Experimental Setup and Design 2. Feedback Process 3. Calculation of Overall Probability of Failure for a Project Team 4. Feedback Statement Development and Rules 5. Feedback Effectiveness Evaluation
Chapter 6	Comparison of Crowd Signals between Experiments I and II	
Chapter 7	Conclusion and Future Work	

Figure 4: Dissertation outline and contents.

2. EXPERIMENT I: CROWD-BASED RISK ASSESSMENT IN STUDENT PROJECTS

In this chapter, I describe the first experiment to collect crowd signals from the student teams and project failure data from the instructors. I used the collected data to compare the occurrence of failure causes between project-based learning and industry (see Chapter 3). I also used the occurrences of project failures, together with the crowd signals, to train the family of failure prediction models (see Chapter 4). The experimental procedures were approved as an exempt study by Purdue's Institutional Review Board (IRB) with protocol #1803020344 and title “Wisdom-of-the-Crowd Signals in Student Engineering Projects” in 2018. The data for the first experiment were collected during the academic year 2018–19. Chapter 2 is organized as follows: Section 2.1 provides a description of the experimental setup and design. Sections 2.2 and 2.3 introduce the crowd signals and instructor questions.

2.1 Experimental Setup and Design

During the recruiting process, I asked the students of design-based courses to volunteer as respondents to a brief survey at the end of each week, answering a set of questions (student crowd signals). The criterion for student recruitment in the study was to be enrolled in a senior-level engineering course that includes a team design project. I visited the courses in person to briefly present the research purposes to the students directly, while the instructor was absent from the classroom. As part of the recruiting process, I created a flyer which included a QR-code and link to the weekly survey (Figure 5).

I provided a gift card incentive for the students. The number of gifts cards and value of each changed per semester: initially (Summer 2018) I included two gift cards of \$50 that were distributed at the end of the semester to two students that responded every week in the semester. In subsequent semesters (Fall 2018/Spring 2019) I switched to a weekly \$20 gift card model because I wanted to encourage participation every week. In communication with the instructors, I also sent some reminders to encourage participation.

Apart from the students, I also asked the instructors of each course to respond to a separate survey at the end of each week, to determine the progress of each student project. The criterion for instructor recruitment was to monitor student teams closely, so they were able to accurately provide the progress of each project.

For confidentiality purposes, the survey was distributed via an anonymous link through Qualtrics, which ensured no student identifiable information was obtained or stored. Instead, I followed an approach where each student used a username of their choice when responding to the survey for the duration of the semester. I informed the students that they should not use their Purdue career account as their username or a username that could make them identifiable to the researchers in any way. The instructors were willingly identified when they agreed to allow me to collect data in their course, but none of their identifiable information was saved in any database.


Wisdom-of-the-Crowd Signals in Student Engineering Projects

Have You Ever Wondered Why
Your Team Project Did Not Turn
Out So Well?


We did too... and with *your* help we can find the answers and improve our understanding of systems engineering failures.

You can be a part of this!
We ask you to respond to our survey questions about various aspects of the student project you are part of.

Are my responses confidential?
Absolutely! When you participate, pick a username you can remember, and use it for the entire semester.



OK, take me to the
survey



QR code and web link
to survey)

ANSWER SURVEY
QUESTIONS ABOUT
YOUR PROJECT

TIME COMMITMENT:
15 MINS PER WEEK

WE WILL GIVE X \$Y GIFT
CARDS AT THE END OF
THE SEMESTER

HAVE QUESTIONS?
CONTACT US!

Karen Marais
AAE Associate Professor
kmarais@purdue.edu

Georgios Georgalis
PhD Student
ggeorgal@purdue.edu

Figure 5: Recruitment flyer as distributed to the students during the recruitment process of the first experiment. The number of gifts cards (X) and value of each (Y) changed per semester: initially (Summer 2018) I included two gift cards of \$50 that were distributed at the end of the semester once. In subsequent semesters (Fall 2018/Spring 2019) I switched to a weekly \$20 gift card model because I wanted to encourage participation every week.

In total, I collected data from 28 different design project teams. The student teams were enrolled in two different courses at Purdue University. All data collection occurred during the academic year 2018–2019. The 28 student teams worked on 18 projects. Some projects spanned multiple semesters, but there was a new student team taking over each semester, so I considered these student teams separately. I collected 240 observations from the two instructors, and 304 observations from the 74 students that participated. For these courses, the student teams worked on either low-gravity fluid experiments or aircraft designs.

Table 3 shows a summary of the number of observations and projects per semester.

Table 3: Summary of data collection during academic year 2018–19. Student teams typically included 4-6 team members. The projects included both hardware and software deliverables as well as progress and final reports.

<i>Semester</i>	<i># of projects</i>	<i>Duration in weeks</i>	<i># student observations</i>	<i># instructor observations</i>
Summer 2018	6	6	56	36 (6 projects for 6 weeks)
Fall 2018	12 (8 new + 4 prev.)	12	218	144 (12 projects for 12 weeks)
Spring 2019	10 (4 new + 6 prev.)	6	30	60 (10 projects for 6 weeks)

2.2 Student Crowd Signals

The crowd signals collect human-centric information (e.g., actions, behaviors, and habits) during the project. Such information may correlate to individual or team performance and therefore to project failures and failure causes as discussed previously in Figure 2. To arrive at a successful set of crowd signals, I surveyed literature that included factors that affect team, project, and individual performance. I included a wide range of literature from the following research areas in the search: human factors, systems engineering, project management, engineering education, psychology, and social sciences. Each factor I identified, led to one or more student questions that applied specifically to the specialized context of student projects. When possible, I phrased the questions so they are hard-to-game, meaning they did not have obvious “correct” answers. I should note that the set of questions I developed are just one way of identifying the presence of a corresponding factor. I also included nine indirect questions that were directly related to project outputs and habits

that may affect time allocation to project work, which have not been studied in previous research work.

In summary, the questions I developed were in the following eight categories:

1. **9 Performance questions (Q1–Q9)** from factors that relate to team performance and/or project success, as identified by human factors, engineering education, and systems engineering literature.
2. **5 Critical Success Factors questions (Q10–Q14)** from the “critical success factors” as identified from project management literature.
3. **5 Individual Personality questions (Q15–Q19)** that include individual personality characteristics that affect team performance as identified from social sciences and psychology literature.
4. **6 Student Estimation questions (Q20–Q25)** that include the students’ own estimations of the project performance.
5. **4 Safety Archetypes questions (Q26–Q29)** that include organizational safety archetypes which relate to dysfunctional team practices that may lead to failures.
6. **9 Indirect signals questions (Q30–Q38)** that include indirect phenomena or habits that may relate to project outcome.
7. **2 Risk Perception questions (Q39–Q40)** that include current risk perception of the team members and may relate to current project status.
8. **9 Individual Actions & Decisions questions (Q41–Q49)** that include cognitive biases of the team members that may show as tendency to particular actions or decisions.

To demonstrate the development process, I describe two examples of hard-to-game questions.

Proactivity is a factor that is associated with project performance because proactive people are willing to take action to affect their environment, in contrast to non-proactive individuals who are less likely to act [Kirkman and Rosen, 1999]. Rather than asking students directly whether they think they are proactive (where the answer would most likely be “yes”), I asked “During the past week, how many times did you attempt to get involved with a project-related task that was outside your immediate responsibility?” (Q2).

The *bandwagon effect* is a cognitive bias where people do or believe things because many other people do or believe the same. Rather than asking members directly whether everyone does or believes the same, I asked “During the past week, did you have any arguments with your team

about the next project actions/tasks?” (Q42). The question is just one way of identifying the presence of the *bandwagon effect*.

Table 4 shows the complete list of the 49 crowd signals, their coded names in the data, and the sources from literature, organized by category.

Table 4: The questions that collected the crowd signals from the students. Each question was based on the definitions of corresponding literature.

<i>Performance</i>			
Q1	Individual Experience (EXP)	The level of proficiency as well as the collective ability to exchange knowledge [Reagans et al., 2005].	How many engineering projects have you participated in so far? Include all engineering projects from coursework, internships, or extracurricular activities. <i>(Integer answer)</i>
Q2	Proactivity (PRO)	Proactive individuals show initiative, are willing to act and affect their environment, and show perseverance [Kirkman and Rosen, 1999].	During the past week, how many times did you attempt to get involved with a project-related task that was outside your immediate responsibility? <i>(Integer answer)</i>
Q3	Stress level (SL)	High level of stress is associated with increased anxiety, negative emotions, distraction, conflict, and loss of team orientation [Dietz et al., 2017].	During the past week, how often were you unable to focus on this project? <i>(Likert scale answer: Never (1) to Always (5))</i>
Q4	Coordination (1) (COO1)	The unification, integration, synchronization of the efforts of group members to provide unity of action in the pursuit of common goals [Salas et al., 2008].	During the past week, how often did you interact with your team members while completing separate project tasks? <i>(Likert scale answer: Never (1) to Always (5))</i>
Q5	Team Impact (IMP)	Teams have been shown to impact the productivity and performance of a project [Hamilton et al., 2003].	During this past week, how often did you think that your team made progress that was meaningful for the success of this project? <i>(Likert scale answer: Never (1) to Always (5))</i>
Q6	Coordination (2) (COO2)	The unification, integration, synchronization of the efforts of group members to provide unity of action in the pursuit of common goals [Salas et al., 2008].	During the past week, for roughly what percentage of your team do you know exactly what they worked on? <i>(Continuous percentage answer)</i>

Table 4 continued

Q7	Standardized work (STND)	Standardized work practices detail how work should be performed [Gilson et al., 2005].	During the past week, rate the level of freedom you felt you had on how to complete your project tasks. (Likert scale answer: No freedom (1) to Complete freedom (5))
Q8	Team Autonomy (AUTO)	High team autonomy has been linked to increased productivity, quality of performance, innovativeness, job satisfaction, decreased turnover, and fewer accidents [van Mierlo et al., 2006; Cordery et al., 2010].	Assume that the course instructor is unavailable for the remaining of the semester. What do you think is the chance your team will successfully complete all the assigned tasks without any oversight for the rest of the semester? (Continuous percentage answer)
Q9	Creativity (CREA)	Teams that explore alternative ways to accomplish their work also should be better able to meet the needs of their customers [Dorst and Cross, 2001; Gilson et al., 2005].	During the past week, which of the following attributes/adjectives relating to creativity do you feel apply to your team's project work? (Multiple answer between 6 adjectives that relate to creativity and 6 that do not)
CSF (Critical Success Factors) [Pinto and Slevin, 1987; Chua et al. 1999]			
Q10	Modularization (MODU)	Modular design, or "modularity in design", is a design approach that subdivides a system into smaller parts called modules or skids, that can be independently created and then used in different systems.	During the past week, roughly what percentage of the tasks you performed could be done independently of the rest of the project? (Continuous percentage answer)
Q11	Clear objectives (COBJ)	To have effective tasks, it is important to plan and pen clearly defined objectives that can deliver desired results.	During the past week, how clearly defined were your team's objectives? (Likert scale answer: Not clear at all (1) to Completely clear (5))
Q12	Commitment (COMT)	The state of being dedicated to a cause.	If your team announced to you today that they all quit, would you be willing to continue working on the project with a completely new team? (Likert scale answer: Definitely not (1) to Definitely yes (5))

Table 4 continued

Q13	Availability of resources (RESO)	Availability means capable of being used or the extent to which resources are available to meet the project's needs.	During the past week, rate your team's availability of resources (tools/space/software/funds) for you to use. (Likert scale answer: Very low availability (1) to Very high availability (5))
Q14	Communication (COMM)	Communication is the act of conveying intended meanings from one entity or group to another through the use of mutually understood signs and semiotic rules.	During the past week, how often did you notice a "silent room" while you were working with your team? (Likert scale answer: Never (1) to Always (5))
Individual Personality [Judge and Bono, 2000; Virgă et al., 2014; Peeters et al., 2006]			
Q15	Neuroticism (NEUR)	Neurotic individuals are associated with low emotional stability, experience frustration, anxiety, depression, and negative emotions.	During the past week, how often did you feel frustrated by your team members or your team's performance? (Likert scale answer: Never (1) to Always (5))
Q16	Openness to experience (OPEN)	Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has.	During the past week, how often did you come up with or agree to a new idea for your project? (Likert scale answer: Never (1) to Always (5))
Q17	Conscientiousness (CONS)	Conscientiousness implies a desire to do a task well, and to take obligations to others seriously. Conscientious people tend to be efficient and organized as opposed to easy-going and disorderly.	During the past week, how often did you skip, delay, postpone, or cancel a task/activity/obligation you were required to do/attend? (Likert scale answer: Never (1) to Always (5))
Q18	Extraversion (EXTR)	Indicates how outgoing and social a person is.	During the past week, how often did you find yourself being the center of the attention of your team? (Likert scale answer: Never (1) to Always (5))
Q19	Agreeableness (AGREE)	Agreeableness manifests itself in individual behavioral characteristics that are perceived as kind, sympathetic, cooperative, warm, and considerate.	During the past week, how often did your team members share detailed about their life with you? (Likert scale answer: Never (1) to Always (5))

Table 4 continued

<i>Student Estimation [adapted from Nolan et al., 2018; Georgalis and Marais, 2019b]</i>			
Q20	Project spending estimate (PROJS)	Students give a qualitative estimate of how much they are spending.	Which of the following reflects your current estimate about your project spending? (Multiple choice: Under/Over/On budget)
Q21	Confidence in project spending estimate (PROJSC)	Confidence in the spending estimate.	How confident are you in your estimate? (Continuous percentage answer)
Q22	Project timeline estimate (PROJT)	Students give a qualitative estimate of whether they are staying on schedule.	Which of the following reflects your current estimate about your project's timeline? (Multiple choice: Ahead of/Behind/On schedule)
Q23	Confidence in project timeline estimate (PROJTC)	Confidence in the timeline estimate.	How confident are you in your estimate? (Continuous percentage answer)
Q24	Project technical performance estimate (PROJP)	Students give a qualitative estimate of whether they are satisfying their requirements.	Which of the following reflects your current estimate about your project's technical performance? (Multiple choice: satisfying fewer/more/as planned requirements)
Q25	Confidence in project technical performance estimate (PROJPC)	Confidence in technical performance estimate.	How confident are you in your estimate? (Continuous percentage answer)
<i>Team Actions & Archetypes [Marais et al., 2006]</i>			
Q26	Unintended side effects of fixes (UNEFF)	Poorly thought-out fixes may have unintended side effects.	If new problems occurred this week, do you think they were handled appropriately? (Multiple choice: Yes/No/Does not apply)
Q27	Stagnant risk management (STRM)	When technological advances are not accompanied by concomitant understanding of the associated risks, risk may increase.	During the past week, did your team consider new potential risks as a result of any new project tasks or updates? (Multiple choice: Yes/No/Does not apply)
Q28	Fixing symptoms rather than root causes (FSYM)	Fixes to problems that only address the symptoms may worsen or prolong the original problem.	During the past week, were you disappointed because a problem that your team thought had been fixed, had instead continued or gotten worse? (Multiple choice: Yes/No/Does not apply)

Table 4 continued

Q29	The vicious cycle of bureaucracy (BUREAU)	When organizations respond to problems with more rules and bureaucracy, employees may become apathetic or alienated.	During the past week, were you frustrated about any rule or bureaucracy that was out of your control? <i>(Multiple choice: Yes/No/Does not apply)</i>
<i>Indirect Signals</i>			
Q30	Number of material outputs (OUTP)	An increase or decrease in hardcopy or electronic files may indicate how much progress the team is making and therefore relate to project performance.	During the past week, did you notice a change in project outputs (hardcopy documents, electronic files, scrap paper to sketch ideas etc.) from your team? <i>(Likert scale answer: Large decrease (1) to Large increase (5))</i>
Q31	Social media engagement (SMENG)	Time spent on social media may be related to distracted individuals are while working on a project.	During the past week, how much time on average per day did you spend on social media platforms? <i>(Multiple choice: <1/ 1-2/ 2-3/ 3-4/ >4 hours)</i>
Q32	Eating habits (1) (EAT1)	Eating habits impact overall individual health and therefore may relate to how individuals perform.	During the past week, which of the following statements best describes your eating habits this week? <i>(Multiple choice: Fast food/ Restaurants/ Home/ Dining Halls)</i>
Q33	Eating habits (2) (EAT2)	Eating habits impact overall individual health and therefore may relate to how individuals perform.	During the past week, did you have breakfast before coming in for class? <i>(Multiple choice: No/Before some/ Before all class times)</i>
Q34	Time spent thinking the project (TSPENT)	How long an individual spends thinking about the project may be correlated to much they contribute to the project.	During the past week, what percent of your working time did you spent thinking about this project or working on this project? <i>(Continuous percentage answer)</i>
Q35	Unscheduled team meetings (TMEET)	Unscheduled team meetings may indicate team effort to meet performance requirements during crunch times.	During the past week, how many times did you and other members of your team arrange to meet and work on the project outside the regular class time? <i>(Integer answer)</i>
Q36	New equipment (NTOOL)	Ordering new supplies may be related to how a project is progressing and are related to project spend.	During the past week, how many items (tools/supplies/project equipment) did your team order? <i>(Integer answer)</i>

Table 4 continued

Q37	Exercising habits (EXERC)	Exercising habits are related to overall individual health and may be related to how individuals perform on a project.	During the past week, how often did you physically exercise? (Multiple choice: 0/ 1-2/ 3-4/>4 times)
Q38	Financial pressure (FPRES)	Financial pressure arises from any situation where money worries are causing stress, which may relate to lack of the individual's focus on a project.	During the past week, how often did you turn down a fun activity because you thought it was too expensive? (Integer answer)
Risk perception [Sjöberg, 1999; Rockenbach et al. 2007]			
Q39	Risk perception (RPERC)	Students rank three hypothetical scenarios from the one they consider the highest risk to the one they consider the lowest risk. The scenarios are related to a cost, schedule, or requirements mishap.	Which of the following events do you consider the highest risk for your project's overall success? (Ranking between a cost/schedule/requirements risk)
Q40	Outcome preference (OUTP)	In the scenario that a failure is bound to happen, students provide the one they think would have the lowest impact.	If you had to choose one of the following failures for your project at the end of the semester, which would have the lowest impact? (Multiple choice cost/schedule/requirements failure)
Individual Actions & Decisions [Lehner et al., 1997; Montibeller and Winterfeldt, 2015; Baybutt, 2018]			
Q41	Ambiguity effect (AMBI)	The tendency to avoid options for which missing information makes the probability seem "unknown".	During the past week, did you disagree with an idea or decision because you thought you did not understand all potential implications? (Multiple choice: Yes/No/Does not apply)
Q42	Bandwagon effect (BANDW)	Tendency to do or believe what others do or believe. As more people come to believe in something, others do too, regardless of the underlying evidence.	During the past week, did you have any arguments with your team about the next project actions/tasks? (Multiple choice: Yes/No/Does not apply)
Q43	Focusing effect (FOCUS)	The tendency to place too much importance on one aspect of an event.	During the past week, can you single out one project decision by your team as the most important? (Multiple choice: Yes/No)

Table 4 continued

Q44	Normalcy bias (NORM)	The refusal to plan for, or react to, a disaster which has never happened before.	During the past week, did you spend any time thinking about how things might go wrong for this project? (Multiple choice: Yes/No)
Q45	Not invented here (NIH)	Aversion to contact with or use of products, research, standards, or knowledge developed outside a group.	During the past week, did you get any new ideas about your project from other teams or people? (Multiple choice: Yes/No)
Q46	Confirmation bias (CONF)	The tendency to search for, interpret, focus on, and remember information in a way that confirms one's preconceptions.	During the past week, did you learn any new things that surprised you, because of your involvement with this project? (Multiple choice: Yes/No/Does not apply)
Q47	Parkinson's Law of Triviality (PARKL)	The tendency to give disproportionate weight to trivial issues.	During the past week, did your team spend significant time discussing what you thought as trivial matters about the project? (Multiple choice: Yes/No/Does not apply)
Q48	Anchoring (ANCHOR)	The tendency to rely too heavily, or "anchor", on one trait or piece of information when making decisions (usually the first piece of information acquired on that subject).	For any new project decisions that you had to make this week, did you think through all viable solutions or go with the one that you thought of first? (Multiple choice: Think through/First thought/Does not apply)
Q49	Overconfidence effect (OVERC)	Excessive confidence in one's own answers to questions.	How confident do you feel about the accuracy of your answers to this questionnaire? (Continuous percentage answer)

2.3 Instructor Questions

The instructors provided answers to a total of 14 questions, as shown in Table 5. Three of the questions captured failure in terms of three project metrics: budget, schedule, and technical performance. Ten questions captured whether a student team showed signs of any of the ten failure causes I considered. Lastly, there was one question to rate each team's productivity for the week. I originally intended for the productivity measure to be part of the crowd signals, as it is an important factor in team performance [Hamilton et al., 2003], but I was unable to find a way to collect the information in an unbiased way from the students. Therefore, I elected for the instructor to provide this productivity evaluation. I treat the productivity measure as a crowd signal for

purposes of future failure prediction as discussed later in Chapter 4.1. I did not consider the productivity measure in the failure cause correlations, as the goal was to identify how the student responses and actions impact the failure cause occurrences, without including any external measures.

Table 5: The questions to the instructors. Three questions captured occurrences of project failures and ten questions captured occurrences of failure causes.

<i>Project Failures</i>		
I1	Budget status	What is currently true about the project budget, compared to what you initially planned? (Multiple choice: Under/On/Over budget)
I2	Schedule status	What is currently true about the project schedule, compared to what you initially planned? (Multiple choice: Ahead of/On/Behind schedule)
I3	Technical requirements status	What is currently true about meeting the technical requirements for the project, compared to what you initially planned? (Multiple choice: Meeting fewer/as planned/more requirements)
<i>Productivity [Hamilton et al., 2003]</i>		
I4	Productivity	Rate each team's productivity. (Likert scale answer: Not productive at all (1) to Extremely productive (5))
<i>Failure causes [Sorenson and Marais, 2016]</i> <i>(Binary choice for each team: Occurrence/Not occurrence)</i>		
I5	Indicate whether a team "Failed to consider an aspect in the system design" this past week.	
I6	Indicate whether a team "Made a decision or action that was not well justified" this past week.	
I7	Indicate whether a team "Did not consider redundant components or measures for their actions" this past week.	
I8	Indicate whether a team "Made a mistake because members lack experience" this past week.	
I9	Indicate whether a team "Did not properly document their progress" this past week.	
I10	Indicate whether a team "Run into communication issues" this past week.	
I11	Indicate whether a team "Did not run adequate tests for their equipment" this past week.	

Table 5 continued

I12	Indicate whether a team “Managed risk poorly” this past week.
I13	Indicate whether a team “Violated rules or procedures” this past week.
I14	Indicate whether a team “Rushed into action without fully understanding the impacts to the system” this past week.

Regarding the budget and schedule metrics from the instructors, I classified any project that is not progressing as planned as a failure for that metric in the given week since there was a divergence from the initial project plan. Regarding the technical requirements metric, if a team is satisfying fewer requirements than planned, I considered it a failure.

I used the instructor data in a variety of ways, depending on the goal. I used the observed instances of failure causes (I5-I14) to compare their occurrence with industry (see Chapter 3.1) and their correlations with the crowd signals to develop the feedback statements (see Chapter 5.4). I used the observed project failures (collected via questions I1-I3), the productivity measure (I4), and crowd signals from the students to train the predictive models for future failures in student projects (see Chapter 4.1).

3. EVALUATION OF PROJECT-BASED LEARNING (PBL) AND RECOMMENDATIONS FOR IMPROVEMENT¹

In this chapter, I evaluate the effectiveness of PBL by questioning how well it prepares students for a frequent engineering phenomenon in professional practice—failure. *Does PBL provide sufficient opportunities for students to fail safely, and learn from the experience? How can we improve?* To answer the question, I compared failure cause occurrence rates from industry, found in literature, with the failure cause occurrence rates from the student teams in Experiment I. I examined ten failure causes as shown earlier in Table 2. The goal was to identify which of the ten failure causes are underrepresented in PBL compared to industry, and therefore areas that instructors can have more impact by making improvements.

Using crowd signals from the students and failure cause occurrences from the instructors (Experiment I), I built logistic regression models to find correlations between specific crowd signals and the occurrence of the underrepresented failure causes. By interpreting the regression coefficients, I suggested specific improvements that instructors could use to give their students more opportunities to learn from specific failure causes during PBL.

Chapter 3 is organized as follows: Section 3.1 provides the statistical result of comparing failure cause occurrences between industry and PBL and identifying the underrepresented failure causes. Section 3.2 focuses on suggestions for PBL improvements to instructors, based on regression models of the crowd signals with the underrepresented failure causes.

3.1 Failure Cause Occurrence: Comparing with Industry

To compare PBL with industry, I identified which of the ten failure causes are underrepresented in student projects by comparing the occurrence rate of the failure causes between two samples: the student (“PBL”) and industry (“IND”) projects.

¹A preliminary version of the research and results in this chapter were originally published in: Georgalis, G. and Marais, K. (2019a) ‘Assessment of Project-Based Learning Courses using Crowd Signals’, *In ASEE 2019 Annual Conference & Exposition*, Tampa, FL.

The null hypothesis was that the occurrence rate of the failure causes in student projects was greater than or equal to the equivalent rate for the industry projects. The null hypothesis was based on two assumptions. First, I assumed that PBL-inspired projects are successful from an educational perspective (i.e., offer sufficient opportunities for the students to experience a particular failure cause before they graduate and join the workforce). Second, I assumed that failure causes are more likely to occur in amateur teams, such as student teams, compared to professional teams. The end goal is to evaluate whether PBL provides sufficient exposure in areas that are useful for students to experience before they join the workforce, and not necessarily to “match” the failure cause occurrences between PBL and industry teams.

To test the hypothesis, I used Barnard’s exact statistical test [Barnard, 1945], which can handle sample proportions. Barnard’s statistical test is a proportion test that has more power than other exact tests and is more accurate for small sample sizes than a chi-squared test [Röhmel and Mansmann, 1999; Suissa and Shuster, 1985]. The test assumes that the two samples “PBL” and “IND” are binomial experiments, which means that each student project and industry project is independently equally likely to show signs of a failure cause. The “PBL” sample included the occurrences of failure causes from the 28 student teams I observed over one academic year (Experiment I). The “IND” sample included failure cause data from 32 industry project failures from literature [Sorenson and Marais, 2016; Aloisio, 2019] that identified and categorized occurrences of failure causes on various non-accident project failures.

To conduct the exact test, I created an estimated *occurrence* measure for both samples. The measure describes the proportion of student and industry projects that included a particular failure cause i . The quality of the PBL occurrence measures is based on the ability of the instructor to identify the failure causes and report them accordingly.

I computed the *occurrence* measure separately for the two samples. For the industry sample, the estimated occurrence is:

$$\hat{O}_{(IND)i} = \frac{\sum_{j=1}^{n_1} TRUE_{(IND)i,j} | Failure}{n_1} \quad (Equation 1)$$

Where i is one of the ten failure causes, n_1 is the number of industry project failures equal to 32, and $TRUE_{(IND)i,j}|Failure$ is a binary variable that is equal to 1 if failure cause i occurred during industry project failure j or 0 if not.

For the “PBL” sample, the estimated *occurrence* for PBL is:

$$\hat{O}_{(PBL)i} = \frac{\sum_{j=1}^{n_2} TRUE_{(PBL)i,j}|Failure}{n_2} \quad (Equation 2)$$

Where i is one of the ten failure causes, n_2 is the number of student projects equal to 28, and $TRUE_{(PBL)i,j}|Failure$ is a binary variable that is equal to 1 if failure cause i occurred during student project j or 0 if not, conditioned on the instructor also observing a project failure in the same week as failure cause i .

With the estimated *occurrence* measures defined for the two binomial samples “PBL” and “industry”, I conducted Barnard’s exact statistical test with the null hypothesis that the actual failure cause occurrence rate $O_{PBL,i}$ in student projects is equal to or greater than the actual occurrence rate $O_{IND,i}$ in industry projects. I rejected the null hypothesis for any $p_i \leq \frac{\alpha}{m} = \frac{0.05}{10} = 0.005$ (Bonferroni correction for ten comparisons).

$$\begin{aligned} H_0: O_{PBL,i} &\geq O_{IND,i} \\ H_a: O_{PBL,i} &< O_{IND,i} \end{aligned} \quad (Equation 3)$$

To conduct the test, I first created a 2x2 contingency table for each failure cause i (the contingency tables for each failure cause are included in Appendix B):

Table 6: Generic contingency table for each of the ten failure causes i .

Failure cause i	“PBL” sample	“IND” sample	Total
<i>Occurrence</i>	$x_{PBL,i}$	$x_{IND,i}$	$x_{PBL,i} + x_{IND,i}$
<i>Not occurrence</i>	$28 - x_{PBL,i}$	$32 - x_{IND,i}$	$60 - x_{PBL,i} - x_{IND,i}$
Total	28	32	60

Based on the previous definitions, the estimated *occurrence* measure for both samples is:

$$\hat{O}_{(PBL)i} = \frac{x_{1,i}}{28} \quad (\text{Equation 4})$$

$$\hat{O}_{(IND)i} = \frac{x_{2,i}}{32} \quad (\text{Equation 5})$$

Under the null hypothesis, the common probability responding to the two groups is p . Then, the probability of obtaining the contingency table, M_o , is the product of two binomials:

$$P(M_o|p) = \binom{28}{x_{1,i}} \binom{32}{x_{2,i}} p^{x_{1,i}+x_{2,i}} (1-p)^{60-x_{1,i}-x_{2,i}} \quad (\text{Equation 6})$$

To obtain the p-value for the test, I considered the critical region that contains all contingency tables that represent an outcome at least as extreme the observed table. Then, the significance level $\alpha(p)$ can be obtained by summing the above probabilities over all the tables in the critical region (CR). The p-value is obtained by maximizing the significance level function $\alpha(p)$ [Mato and Andrés, 1997]:

$$\alpha(p) = \sum_{CR} P(M_o|p) \quad (\text{Equation 7})$$

$$\alpha^* = \max_{0 < p < 1} (\alpha(p)) \quad (\text{Equation 8})$$

After the completion of the statistical test for all ten failure causes, I identified which of the failure causes are underrepresented in PBL compared to industry projects (i.e., the ones that the null hypothesis is rejected with a p-value less than the 0.005). Table 7 shows the results of the statistical test.

Table 7: Occurrence measures and Barnard's statistical test results for the failure causes. 1 out of 10 failure causes are underrepresented in the 28 student projects I studied compared to the 32 industry projects from [Sorenson and Marais, 2016 and Aloisio, 2019].

<i>Failure cause</i>	$x_{1,i}$	$x_{2,i}$	$\hat{O}_{(PBL)i}$	$\hat{O}_{(IND)i}$	<i>Barnard's test one-tailed p-value</i>	<i>H0 rejected?</i>
Failed to consider design aspect	10	29	35.7	90.6	0.000002811	Yes
Used inadequate justification	9	11	32.1	34.4	0.456342494	No
Failed to form a contingency plan	7	8	25.0	25.0	1.000000000	No
Lacked experience	8	15	28.6	46.9	0.096511165	No
Kept poor records	5	4	17.9	12.5	0.655108668	No
Inadequately communicated	6	11	21.4	34.3	0.148450314	No
Subjected to inadequate testing	5	15	17.9	46.9	0.009647157	No
Managed risk poorly	5	12	17.9	37.5	0.054538892	No
Violated procedures	3	5	10.7	15.6	0.347870305	No
Did not allow system aspect to stabilize	6	16	21.4	50.0	0.012635167	No

The statistical test shows that of the 10 failure causes, *failed to consider a design aspect* statistically appears less frequently in the student projects compared to industry. This failure cause is fundamentally linked to proper systems engineering: it represents a scenario where for example, the students failed to consider an important requirement during system development, or did not detect a design flaw, or component incompatibility.

3.2 PBL Improvement Recommendations

After I identified *failure to consider a design aspect* (FC1, Table 2) as underrepresented in PBL, I used logistic regression to find the correlations between the crowd signals and the occurrence of FC1 and used that information as a guide to suggest improvements to instructors. The goal is for the improvements to provide students with more experiences of missing a design aspect, to learn to overcome this type of failure cause before they join the workforce.

I followed a 3-step process:

1. I identified which of the student questions are actionable from the instructor (i.e., an instructor action may impact how students respond to that particular question).

2. I found the correlations of each of the crowd signals with FC1 *failure to consider a design aspect*.
3. I used the correlations of the actionable crowd signals to suggest improvements.

Even if the crowd signals are collected from the students, the instructors do have a direct or indirect capability to alter the course environment in such a way that the student experience is different (and as a result the student answers to our questions). Therefore, as a first step I start by identifying which of the crowd signals can be influenced by the instructor.

For example, Q1: “How many engineering projects have you participated in so far? Include all engineering projects from coursework, internships, or extracurricular activities.”, is a crowd signal that the instructor may influence directly. The instructor could split students into teams based on their experience level or could make the course available only to students with a particular prerequisite that provides the necessary experience, or change the student class the course is offered at. Table 8 summarizes the student questions that the instructor can influence and the justification for each, organized by question category.

Table 8: The student questions that are actionable from the instructor with associated justification.

<i>Q*</i>	<i>Question</i>	<i>Justification</i>
Q1	How many engineering projects have you participated in so far? Include all engineering projects from coursework, internships, or extracurricular activities. (<i>Integer answer</i>)	The instructor can change how teams are formed based on experience, allow students to enroll with pre-requisites, or change which student class the course is offered at.
Q4	During the past week, how often did you interact with your team members while completing separate project tasks? (<i>Likert scale answer: Never (1) to Always (5)</i>)	The instructor can change the course setting and location to promote or inhibit frequent interaction among team members. For example, some separate tasks may need to be completed at different locations, which makes interacting frequently less likely.
Q6	During the past week, for roughly what percentage of your team do you know exactly what they worked on? (<i>Continuous percentage answer</i>)	The instructor can have an impact on how often and for how long the team meets together. For example, they can have different settings and timelines for different tasks, which makes it hard for team members to know what each other is working on.

Table 8 continued

<i>Q*</i>	<i>Question</i>	<i>Justification</i>
Q7	During the past week, rate the level of freedom you felt you had on how to complete your project tasks. <i>(Likert scale answer: No freedom (1) to Complete freedom (5))</i>	The instructor can directly impact how much freedom the students think they have: an instructor can be more open to ideas outside of what they had thought, or reject any student suggestions.
Q8	Assume that the course instructor is unavailable for the remaining of the semester. What do you think is the chance your team will successfully complete all the assigned tasks without any oversight for the rest of the semester? <i>(Continuous percentage answer)</i>	The instructor can impact the students' autonomy perception by altering the way and type of oversight they provide. An instructor can be very involved with all student decisions and review everything in detail or let the students be more autonomous by proceeding with their own decisions without strict oversight.
Q10	During the past week, roughly what percentage of the tasks you performed could be done independently of the rest of the project? <i>(Continuous percentage answer)</i>	The instructor can directly impact modularity of a project by how they have split up the various tasks.
Q11	During the past week, how clearly defined were your team's objectives? <i>(Likert scale answer: Not clear at all (1) to Completely clear (5))</i>	The instructor can influence how clear the students think their objectives are by changing in what way the objectives are communicated, how specific they are, or how often they change.
Q13	During the past week, rate your team's availability of resources (tools/space/software/funds) for you to use. <i>(Likert scale answer: Very low availability (1) to Very high availability (5))</i>	The instructor can directly change the availability of resources by allowing students to spend more, use additional rooms, computers, or tools.
Q20	Which of the following reflects your current estimate about your project spending? <i>(Multiple choice: Under/Over/On budget)</i>	The instructor may influence what the students think about the current project metrics by giving them feedback about their progress. The level of information the instructor shares on these metrics, may also impact the confidence of the students' answers. For example, the instructor can make a comment about a team being late with their delivery in a particular week, which also makes the students know that they are behind schedule with high confidence.
Q21	How confident are you in your estimate? <i>(Continuous percentage answer)</i>	
Q22	Which of the following reflects your current estimate about your project's timeline? <i>(Multiple choice: Ahead of/Behind/On schedule)</i>	
Q23	How confident are you in your estimate? <i>(Continuous percentage answer)</i>	

Table 8 continued

<i>Q*</i>	<i>Question</i>	<i>Justification</i>
Q24	Which of the following reflects your current estimate about your project's technical performance? (Multiple choice: satisfying fewer/more/as planned requirements)	
Q25	How confident are you in your estimate? (Continuous percentage answer)	
Q27	During the past week, did your team consider new potential risks as a result of any new project tasks or updates? (Multiple choice: Yes/No/Does not apply)	The instructor can actively encourage students to re-evaluate previous risk considerations when they make new updates to their project.
Q29	During the past week, were you frustrated about any rule or bureaucracy that was out of your control? (Multiple choice: Yes/No/Does not apply)	The instructor can have procedures in place that increase the bureaucratic burden on the students (e.g., specific documentation of certain tasks or processing times in ordering of parts)
Q36	During the past week, how many items (tools/supplies/project equipment) did your team order? (Integer answer)	The instructor can give the students more budget allowance in ordering parts, which can directly impact how many new supplies they order.
Q40	If you had to choose one of the following failures for your project at the end of the semester, which would have the lowest impact? (Multiple choice cost/schedule/requirements failure)	The instructor can influence risk perception by verbally putting more emphasis into one outcome over another.
Q45	During the past week, did you get any new ideas about your project from other teams or people? (Multiple choice: Yes/No/Does not apply)	The instructor can impact how often different teams communicate by altering the course environment. For example, multiple teams working in the same room would facilitate the exchange of ideas between teams.
Q46	During the past week, did you learn any new things that surprised you, because of your involvement with this project? (Multiple choice: Yes/No/Does not apply)	The instructor's teaching approach can impact how much a student learns in a course. For example, a course may include lectures and training on top of the design work.

To find the correlations of each of the crowd signals with FC1: *Failure to consider a design aspect*, I used the individual student responses to the 49 questions (crowd signals) as the independent features $X_{i,t}$ of the failure cause FC1, during a project week t . The dependent variable $Y_{FC1,t}$ expresses the binary occurrence of failure cause FC1, i.e., $Y_{FC1,t} = 1$ when FC1 occurred, and $Y_{FC1,t} = 0$ when FC1 did not occur.

For this classification problem (i.e., binary dependent variables: occurrence or not), I used a logistic regression model. I considered and attended to some of the logistic regression assumptions. The data from the crowd signals was in panel form: it included repeated measurements from partly the same students. Regression models are built on the assumption that observations are independent, which does not hold here, as the responses from the same student at different times are not independent. One common way to account for non-independence of panel observations in logistic regression models is to include random effects [Harrison et al., 2018]. Random effects account for non-independence of the multiple responses coming from a single subject and allow estimation of variance between different students. With random effects, each student has their own intercept term in the model. Models with random effects assume that uncontrolled person-specific effects (e.g., age or gender) are not correlated with the predictors. The random effects take a different value for each student i and appear in the model as c_i .

Based on these considerations, the models were of the following form when predicting failure cause FC1:

$$\log\left(\frac{\hat{p}_{FC1,t}(\hat{Y}_{FC1,t} = 1)}{1 - \hat{p}_{FC1,t}(\hat{Y}_{FC1,t} = 1)}\right) = a + bX_{i,t}^T + c_i + \varepsilon_{it} \quad (\text{Equation 9})$$

Where a is the intercept constant, b is a column vector of slopes for each feature, $X_{i,t}^T$ is a row vector of the 49 predictors at week t , $c_i \sim N(0, \sigma_i^2)$ are individual random effects, and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ is the observation-specific random error. $\hat{p}_{FC1,t}(\hat{Y}_{FC1,t} = 1)$ is the probability of failure cause FC1 occurring during week t .

I built the linear mixed effects models using the computational package *lme4* for R [Bates et al., 2007] and *blme* for R, which provides a wrapper to *lme4* by adding the maximum penalized likelihood approach developed by Chung et al. (2013), to account for singularities in complex mixed models. Since the features $X_{i,t}$ are the crowd signals and express different data forms, I used coding schemes and scaling to normalize them. Table 9 summarizes the coding schemes for all crowd signals. For this model, I discarded observations with missing values.

Table 9: Coding schemes for the 49 crowd signals , based on data type.

<i>Data type</i>	<i>Applicable questions</i>	<i>Coding scheme</i>
5-point full Likert scale	Q3, Q4, Q5, Q7, Q11–Q19, Q30	Coded as integer 1–5 for each level
Integer	Q1, Q2, Q35, Q36, Q38	Not coded, treated as integer
Continuous percentage	Q6, Q8, Q10, Q21, Q23, Q25, Q34, Q49	Not coded, treated as continuous
Categorical	Q20, Q22, Q24, Q26–Q29, Q31–Q33, Q37, Q39–Q48	Coded as categorical (using one-hot encoding)
Character multiple answer	Q9	<p>Adjectives associated with creative designs count as +1, and their opposite adjectives as -1. The sum value is then in the range: $-6 \leq \sum adj \leq 6$.</p> <p>Then code the sum as categorical with the balanced scheme:</p> $Creativity = \begin{cases} \text{Low, } \sum adj \leq -2, \\ \text{High, } \sum adj \geq 2 \\ \text{Moderate, otherwise} \end{cases}$

Table 10 shows the model coefficients for failure cause FC1.

The coefficients of the predictor variables from the logistic regression models are interpreted in terms of the log-odds of failure cause. For example, the coefficient for experience $b_1 = -0.086$ is the expected change in the log-odds of failing to consider a design aspect for a one-unit increase in experience, while keeping all other predictors at fixed values. Equivalently, the odds ratio can be calculated by exponentiating the coefficient value to get 0.917 which means we expect to see

about 8.3% decrease in the odds of missing a design aspect, for a one-unit increase in experience, while keeping all other variables at fixed values. The coefficients with p-values < 0.05 are bolded in the resulting model table as the model determines these coefficients are non-zero (i.e., there is enough evidence in the data about the existence of a correlation between the particular crowd signal and the specific failure cause).

The resulting model for FC1 shows that the likelihood of occurrence for this failure cause increases when students respond “No” (relative to “Does not apply”) to whether they were frustrated with rules that they can’t control (Q29, $b_{36} = 3.992$). The result may indicate that the absence of important rules or thorough reporting procedures makes it more likely for a project team to miss an important design aspect.

To the contrary, when students interact frequently with each other when completing independent tasks (Q4, $b_4 = -0.636$), knowing what most of their team worked on (Q6, $b_6 = -0.524$), are not disappointed from previous problems continuing (Q28, $b_{34} = -3.380$), ordering more equipment (Q36, $b_{50} = -2.053$), and not having arguments (Q42, $b_{61} = -2.447$), the likelihood of failing to miss a design aspect decreases. These results all support positive practices for a team project such as increasing communication, interaction, and making progress. The model also shows that when students think new problems are not handled properly (Q26, $b_{30} = -2.331$), the failure cause occurrence likelihood also decreases, possibly hinting that when a team puts some thought and effort into new problems (even if they fail to find a solution), it is less likely to miss a design aspect. Lastly, when students consider a requirements failure as preferable and less impactful for the project than a cost failure (Q40, $b_{57} = -3.093$), the likelihood of *missing a design aspect* also decreases. My interpretation of this correlation was that for the context of short-term student projects, teams may consider a requirements problem fixable, whereas schedule or cost failure would cause permanent problems to the project and perhaps a lower grade. Teams that thought like that, may have generally been more careful about delivering, reducing the likelihood of FC1.

Table 10: Mixed-effects logistic regression model coefficients for FC1: Failure to consider a design aspect.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	-1.770 (2.992)	$b_{30}(\text{Q26} = \text{No})$	-2.331 (1.081)*	$b_{60}(\text{Q41} = \text{Yes})$	1.308 (1.130)
$b_1(\text{Q1})$	-0.086 (0.241)	$b_{31}(\text{Q26} = \text{Yes})$	-0.257 (0.959)	$b_{61}(\text{Q42} = \text{No})$	-2.447 (1.14)*
$b_2(\text{Q2})$	-0.046 (0.233)	$b_{32}(\text{Q27}=\text{No})$	0.154 (1.109)	$b_{62}(\text{Q42} = \text{Yes})$	-2.120 (1.183)^^
$b_3(\text{Q3})$	0.336 (0.236)	$b_{33}(\text{Q27}=\text{Yes})$	0.269 (1.023)	$b_{63}(\text{Q43} = \text{Yes})$	-0.107 (0.578)
$b_4(\text{Q4})$	-0.636 (0.257)*	$b_{34}(\text{Q28}=\text{No})$	-3.380 (1.293)**	$b_{64}(\text{Q44} = \text{Yes})$	-0.178 (0.470)
$b_5(\text{Q5})$	0.395 (0.231)^^	$b_{35}(\text{Q28}=\text{Yes})$	-2.076 (1.349)	$b_{65}(\text{Q45} = \text{Yes})$	0.114 (0.507)
$b_6(\text{Q6})$	-0.524 (0.237)*	$b_{36}(\text{Q29}=\text{No})$	3.992 (1.537)**	$b_{66}(\text{Q46} = \text{No})$	-1.509 (1.214)
$b_7(\text{Q7})$	-0.280 (0.243)	$b_{37}(\text{Q29}=\text{Yes})$	2.918 (1.606)^^	$b_{67}(\text{Q46} = \text{Yes})$	-1.302 (1.194)
$b_8(\text{Q8})$	-0.143 (0.252)	$b_{38}(\text{Q30})$	0.408 (0.216)^^	$b_{68}(\text{Q47} = \text{No})$	1.092 (1.545)
$b_9(\text{Q9} = \text{Low})$	-0.404 (1.781)	$b_{39}(\text{Q31} = 2\text{-}3\text{h})$	-0.658 (0.606)	$b_{69}(\text{Q47} = \text{Yes})$	0.407 (1.602)
$b_{10}(\text{Q9} = \text{Moderate})$	0.269 (0.441)	$b_{40}(\text{Q31} = 3\text{-}4\text{h})$	-0.157 (1.037)	$b_{70}(\text{Q48} = \text{First thought})$	0.373 (0.803)
$b_{11}(\text{Q10})$	0.235 (0.219)	$b_{41}(\text{Q31} = <1\text{h})$	0.021 (0.58)	$b_{71}(\text{Q48} = \text{Think through})$	0.525 (0.704)
$b_{12}(\text{Q11})$	0.052 (0.226)	$b_{42}(\text{Q31} = >4\text{h})$	-0.870 (1.069)	$b_{72}(\text{Q49})$	-0.110 (0.242)
$b_{13}(\text{Q12})$	0.069 (0.238)	$b_{43}(\text{Q32} = \text{Dining hall})$	0.910 (1.029)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.2351 Random effects $c_i \sim N(0, 0.311^2)$ Occurrence ratio in data: 0.211</p>	
$b_{14}(\text{Q13})$	0.233 (0.228)	$b_{44}(\text{Q32} = \text{Restaurants})$	1.473 (0.986)		
$b_{15}(\text{Q14})$	0.010 (0.227)	$b_{45}(\text{Q32} = \text{Home})$	0.115 (0.743)		
$b_{16}(\text{Q15})$	0.111 (0.242)	$b_{46}(\text{Q33}=\text{No})$	0.559 (0.586)		
$b_{17}(\text{Q16})$	0.213 (0.223)	$b_{47}(\text{Q33}=\text{Some})$	0.011 (0.598)		
$b_{18}(\text{Q17})$	0.292 (0.266)	$b_{48}(\text{Q34})$	-0.144 (0.254)		
$b_{19}(\text{Q18})$	-0.247 (0.249)	$b_{49}(\text{Q35})$	0.026 (0.251)		
$b_{20}(\text{Q19})$	-0.382 (0.228)^^	$b_{50}(\text{Q36})$	-2.053 (1.028)*		
$b_{21}(\text{Q20} = \text{Over budget})$	0.303 (0.715)	$b_{51}(\text{Q37} = >3\text{-}4)$	-0.334 (0.64)		
$b_{22}(\text{Q20} = \text{Under budget})$	-0.167 (0.548)	$b_{52}(\text{Q37} = >4)$	-0.139 (0.698)		
$b_{23}(\text{Q21} = \text{Behind sched.})$	1.153 (1.018)	$b_{53}(\text{Q37} = \text{None})$	-0.805 (0.609)		
$b_{24}(\text{Q21} = \text{On sched.})$	1.049 (0.954)	$b_{54}(\text{Q38})$	0.035 (0.248)		
$b_{25}(\text{Q22} = \text{More reqs})$	0.966 (0.831)	$b_{55}(\text{Q39}=\text{Reqs})$	-0.415 (0.660)		
$b_{26}(\text{Q22}=\text{reqs as planned})$	-0.433 (0.585)	$b_{56}(\text{Q39}=\text{Sched})$	-0.035 (0.530)		
$b_{27}(\text{Q23})$	-0.041 (0.261)	$b_{57}(\text{Q40}=\text{Reqs})$	-3.093 (1.578)*		
$b_{28}(\text{Q24})$	0.330 (0.297)	$b_{58}(\text{Q40}=\text{Sched})$	0.250 (0.531)		
$b_{29}(\text{Q25})$	0.094 (0.279)	$b_{59}(\text{Q41} = \text{No})$	1.785 (1.111)		

For the last step, I synthesize the previous information (knowing which crowd signals are actionable and which crowd signals correlate with FC1: Q4, Q6, Q29, Q36, and Q40) to suggest recommendations for the instructors that want to improve PBL. Given that FC1 is underrepresented, the suggestions aim to increase the experience of that failure cause among student teams, to prepare them better for industry, by giving them good chances to learn from it in an educational setting.

Table 11: Instructor recommendations to improve student preparation in dealing with failure cause FC1: “Missing a design aspect” in PBL.

<i>Q*</i>	<i>Correlation to FC1 (from Table 10)</i>	<i>Instructor recommendation</i>
Q4	Negative correlation. More frequent interaction while completing independent project tasks, reduces the likelihood of FC1.	The instructor could arrange the course so that specific tasks happen in settings where students do not interact with each other as much. For example, the manufacturing of a component does not happen in the same room as the electrical wiring testing. In that setting, students are possible to miss on a detail that they then must communicate better to fix, learning that interacting within the team and knowing what each other is working on is key to success.
Q6	Negative correlation. Knowing more about what other team members are working on reduces the likelihood of FC1.	
Q29	Positive correlation. Not being frustrated because of a rule or bureaucracy increases the likelihood of FC1.	I interpreted this correlation to signal that the instructors’ rules or processes (if they exist) are simple and therefore do not cause frustration to the students, but also do not enhance the learning process in any way. The instructor could introduce reporting rules and processes that resemble industry standards, but are simple for the students. Even if students complain about such rules, the learning benefit is likely worth it when they familiarize with these processes.
Q36	Negative correlation. The larger number of ordered parts, the likelihood of FC1 decreases.	The correlation may indicate that student teams with ample resources and lack of constraints when it comes to tool or equipment usage, are less likely to miss a design aspect. The instructor could implement realistic equipment/tool usage and expense constraints to the student teams. From a learning perspective, it is helpful for the students to familiarize with such constraints before joining the workforce.
Q40	Negative correlation. When students consider a requirements failure as lower impact and prefer to miss requirements (compared to a budget failure), the likelihood of FC1 decreases.	The instructor could put more emphasis on the requirements of the project, the importance of them, and how they clearly relate to project success.

4. CROWD-BASED FAILURE PREDICTION IN STUDENT PROJECTS²

Chapter 4 focuses on building the failure prediction models to accomplish process (a) of the prototype (see Figure 3). In total, there are three models, one for each performance metric: budget, schedule, and requirements. I trained the models with the data collected from the students and instructors during Experiment 1.

Chapter 4 is organized as follows: Section 4.1 discusses the training process for the three prediction models and associated conclusions. Section 4.2 is an evaluation of the initial models using cross-validation. Section 4.3 applies stepwise reduction and a best subsets approach to reduce the model inputs and improve prediction accuracy.

4.1 Prediction Model Training

To accomplish the goal of predicting future project failures in student projects, I used the following information from a given project week t :

1. the individual student responses to the 49 questions (crowd signals),
2. the current state of the project as provided by the instructor, and
3. the current productivity measure of the team as provided by the instructor,

as the independent predictors $X_{i,t}$ of a failure j during the next project week $t+1$. There are three types of failures: $j=1$ corresponds to cost failure, $j=2$ corresponds to schedule failure, and $j=3$ corresponds to requirements failure. The dependent variable $Y_{j,t+1}$ expresses the binary occurrence of the three failure types, i.e., $Y_{j,t+1} = 1$ when a failure occurs, and $Y_{j,t+1} = 0$ when a failure does not occur.

For this classification problem, I follow a similar mixed effects logistic regression process as discussed previously in Section 3.2. The main difference in this formulation is the inclusion of the two additional instructor inputs at week t (current state of project and productivity) as predictors,

²Part of the research and results in this chapter were originally published in: Georgalis, G. and Marais, K. (2021) ‘Predicting failure events from crowd-derived inputs: schedule slips and missed requirements’, *In INCOSE International Symposium*, Vol. 31, No. 1.

and that the goal of these models is prediction of the project state at week $t+1$. I included these two additional inputs due to the justification from industry that project performance history impacts future performance (i.e., a project that is behind schedule for a few weeks is more likely to stay behind schedule compared to a project that is on schedule), and because team productivity is a factor on team performance [Hamilton et al., 2003].

Based on these considerations, the models are of the following form when predicting the three types of failure ($j = 1, 2, 3$):

$$\log \left(\frac{\hat{p}_{j,t+1}(\hat{Y}_{j,t+1} = 1)}{1 - \hat{p}_{j,t+1}(\hat{Y}_{j,t+1} = 1)} \right) = a + bX_{i,t}^T + c_i + \varepsilon_{it} \quad (\text{Equation 10})$$

Where a is the intercept constant, b is a column vector of the slopes for each predictor, $X_{i,t}^T$ is a row vector of the 51 predictors at week t , $c_i \sim N(0, \sigma_i^2)$ are individual random effects, and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ is the observation-specific random error. $\hat{p}_{j,t+1}$ is the probability of failure j occurring during week $t+1$.

Table 12 summarizes the variables for the three failure prediction models.

Table 12: Predictors and dependent variables for student project failure prediction. I built three models (one for each failure: budget, schedule, and technical requirements), from 51 predictors.

<i>Independent variables (predictors) $X_{i,t}$ at week t</i>		
(1)	Crowd signals	$X_{1-49,t}$ Came from the student responses to Q1–Q49
(2)	Current state of the project	$X_{50,t}$ Came from the instructor's response to I1–I3 (see Table 5), depending on the metric
(3)	Productivity of the team	$X_{51,t}$ Came from the instructor's response to I4 (see Table 5).
<i>Dependent variable $Y_{j,t+1}$ at week $t+1$</i>		
Predicting failure in terms of metric j at week $t+1$	$Y_{j,t+1}$	Came from the instructors' response to I1–I3 from the following week. j = 1 corresponds to the budget metric j = 2 corresponds to the schedule metric j = 3 corresponds to the technical performance metric

I coded the input crowd signals for the prediction models as described before for the failure cause correlation model (see Table 9). In addition to the crowd signals, I also coded the additional instructor measures as shown in Table 13.

Table 13: Coding schemes for the input instructor measures, based on data type.

<i>Data type</i>	<i>Applicable questions</i>	<i>Coding scheme</i>
5-point full Likert scale	I4	Coded as integer 1–5 for each level
Categorical	I1–I3	Coded as categorical (using one-hot encoding)

The following tables show the predictive models for the three types of project failures. The coefficients with p-values < 0.05 are bolded in the tables. The requirements model shows a larger variance in the random effects compared to the other two models.

The model that predicts budget failure indicates that when students perceive they have increased freedom on what to do with the project (Q7, $b_7 = -0.843$), think they are satisfying requirements more requirements planned (Q22, $b_{25} = -1.755$), and do not have problems continue or become worse (Q28, $b_{34} = -3.716$), the likelihood of a cost failure reduces. In contrast, budget failure likelihood increases when students perceive a schedule failure as higher risk compared to a cost failure (Q39, $b_{56} = 1.716$), when they disagree because of lack of understanding decision implications (Q41, $b_{60} = 2.711$), when they single out a decision as most important (Q43, $b_{63} = 1.815$), and when having a budget failure the previous week (I1, $b_{73} = 1.225$).

The model that predicts schedule failure indicates that when students are sharing about their lives (Q19, $b_{20} = -0.538$), think they are spending more funds than they should (Q20, $b_{21} = -1.949$), are turning down activities that they consider fun (Q38, $b_{54} = -0.784$), and understand all potential implications of an action (Q41, $b_{59} = -2.202$), a schedule failure is less likely. For Q48, both the “yes” and “no” options correlate to lower likelihood of a schedule failure compared to the “Do not apply” option. Coming up with solutions either by looking through many options or considering only the first one is better than coming up with no solutions. Going with the first

solution saves time and so for avoiding schedule failure is a better option (larger negative coefficient). On the contrary, with increasing student confidence in their success without oversight (Q8, $b_8 = 0.726$), coming up with or agreeing to more project ideas (Q16, $b_{17} = 0.483$), thinking they are satisfying requirements as planned (Q22, $b_{26} = 1.551$), having previous problems resurface due to poor previous solutions (Q28, $b_{35} = 3.469$), spending more than 4 hours on social media per day (Q31, $b_{39} = 1.28$), perceiving schedule as the highest risk for the project (Q39, $b_{56} = 2.21$), not learning any new things (Q46, $b_{66} = 4.459$), and having a schedule failure the previous week (I2, $b_{73} = 1.286$), all increase the likelihood of a schedule failure.

The model that predicts failure regarding the technical requirements indicates that not exercising at all during the week (Q37, $b_{53} = -2.405$), discussing trivial matters during the project (Q47, $b_{68} = -3.827$) and being increasingly confident in one's answers to the questions (Q49, $b_{72} = -0.759$) reduce the likelihood of a failure. In contrast, when students are increasingly unable to focus on the project (Q3, $b_3 = 0.907$), show low creativity (Q9, $b_9 = 3.683$), introduce new ideas to the project (Q16, $b_{17} = 0.667$), skip or postpone required tasks (Q17, $b_{18} = 0.906$), think they are spending less than they should (Q20, $b_{22} = 1.773$), report their cost estimate with high confidence (Q23, $b_{27} = 0.870$), and having a requirements failure the previous week (I3, $b_{73} = 2.186$), the likelihood of a future requirements failure increases.

Table 14: Mixed-effects logistic regression model for prediction of budget failure.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	-5.471 (3.418)	b_{30} (Q26 = No)	0.077 (1.361)	b_{60} (Q41 = Yes)	2.711 (1.373)*
b_1 (Q1)	0.060 (0.271)	b_{31} (Q26 = Yes)	0.986 (1.345)	b_{61} (Q42 = No)	-0.927 (1.307)
b_2 (Q2)	-0.173 (0.374)	b_{32} (Q27=No)	-0.544 (1.148)	b_{62} (Q42 = Yes)	-0.327 (1.394)
b_3 (Q3)	0.512 (0.282)^	b_{33} (Q27=Yes)	-0.259 (1.091)	b_{63} (Q43 = Yes)	1.815 (0.727)*
b_4 (Q4)	-0.5 (0.284)^	b_{34}(Q28=No)	-3.716 (1.463)*	b_{64} (Q44 = Yes)	0.411 (0.571)
b_5 (Q5)	0.067 (0.267)	b_{35} (Q28=Yes)	-2.668 (1.406)^	b_{65} (Q45 = Yes)	-0.724 (0.554)
b_6 (Q6)	0.109 (0.275)	b_{36} (Q29=No)	1.336 (1.424)	b_{66} (Q46 = No)	0.690 (1.460)
b_7(Q7)	-0.843 (0.296)**	b_{37} (Q29=Yes)	2.471 (1.482)^	b_{67} (Q46 = Yes)	0.359 (1.441)
b_8 (Q8)	-0.177 (0.31)	b_{38} (Q30)	0.424 (0.275)	b_{68} (Q47 = No)	1.652 (2.052)
b_9 (Q9 = Low)	3.341 (1.76)^	b_{39} (Q31 =2-3h)	-0.629 (0.684)	b_{69} (Q47 = Yes)	0.221 (2.086)
b_{10} (Q9 = Moderate)	0.906 (0.526)^	b_{40} (Q31 = 3-4h)	-0.677 (1.344)	b_{70} (Q48 = First thought)	-1.411 (0.995)
b_{11} (Q10)	-0.24 (0.258)	b_{41} (Q31 = <1h)	-1.282 (0.700)^	b_{71} (Q48 = Think through)	-1.064 (0.793)
b_{12} (Q11)	0.012 (0.258)	b_{42} (Q31 = >4h)	-0.663 (1.102)	b_{72} (Q49)	0.252 (0.259)
b_{13} (Q12)	0.005 (0.278)	b_{43} (Q32 = Dining hall)	1.132 (1.177)	b_{73}(I1 = Failure)	1.225 (0.574)*
b_{14} (Q13)	0.092 (0.247)	b_{44} (Q32 = Restaurants)	1.403 (1.335)	b_{74} (I4)	-0.519 (0.268)^
b_{15} (Q14)	0.024 (0.287)	b_{45} (Q32 = Home)	-0.392 (0.943)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.1273 Random effects $c_i \sim N(0, 0.567^2)$ Occurrence ratio in data: 0.254</p>	
b_{16} (Q15)	-0.387 (0.301)	b_{46} (Q33=No)	0.024 (0.722)		
b_{17} (Q16)	-0.375 (0.251)	b_{47} (Q33=Some)	-0.084 (0.663)		
b_{18} (Q17)	0.434 (0.317)	b_{48} (Q34)	-0.170 (0.295)		
b_{19} (Q18)	-0.072 (0.276)	b_{49} (Q35)	-0.435 (0.308)		
b_{20} (Q19)	0.119 (0.264)	b_{50} (Q36)	-1.860 (1.251)		
b_{21} (Q20 = Over budget)	-2.045 (1.066)^	b_{51} (Q37 = 3-4)	-0.241 (0.714)		
b_{22} (Q20 = Under budget)	0.020 (0.626)	b_{52} (Q37 = >4)	0.246 (0.774)		
b_{23} (Q21 = Behind sched.)	0.267 (1.213)	b_{53} (Q37 = None)	0.668 (0.739)		
b_{24} (Q21 = On sched.)	0.976 (1.098)	b_{54} (Q38)	-0.392 (0.28)		
b_{25}(Q22 = More reqs)	-1.755 (0.875)*	b_{55} (Q39=Reqs)	0.196 (0.823)		
b_{26} (Q22=reqs as planned)	1.422 (0.836)^	b_{56}(Q39=Sched)	1.716 (0.652)**		
b_{27} (Q23)	0.515 (0.315)	b_{57} (Q40=Reqs)	-1.863 (1.477)		
b_{28} (Q24)	0.198 (0.32)	b_{58} (Q40=Sched)	0.279 (0.599)		
b_{29} (Q25)	0.394 (0.311)	b_{59} (Q41 = No)	1.815 (1.25)		

Table 15: Mixed-effects logistic regression model for prediction of schedule failure.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	-3.732(2.927)	b_{30} (Q26 = No)	1.18(1.089)	b_{60} (Q41 = Yes)	-0.865(1.058)
b_1 (Q1)	-0.008(0.242)	b_{31} (Q26 = Yes)	1.335(1.003)	b_{61} (Q42 = No)	0.346(1.218)
b_2 (Q2)	0.047(0.205)	b_{32} (Q27=No)	-0.659(1.192)	b_{62} (Q42 = Yes)	-0.611(1.238)
b_3 (Q3)	0.392(0.249)	b_{33} (Q27=Yes)	0.003(1.097)	b_{63} (Q43 = Yes)	-0.921(0.559)^
b_4 (Q4)	-0.387(0.244)	b_{34} (Q28=No)	2.179(1.331)	b_{64} (Q44 = Yes)	0.238(0.518)
b_5 (Q5)	-0.159(0.228)	b_{35}(Q28=Yes)	3.469(1.405)*	b_{65} (Q45 = Yes)	0.845(0.509)^
b_6 (Q6)	0.139(0.231)	b_{36} (Q29=No)	-2.513(1.488)^	b_{66}(Q46 = No)	4.459(1.545)**
b_7 (Q7)	-0.042(0.23)	b_{37} (Q29=Yes)	-1.654(1.455)	b_{67} (Q46 = Yes)	2.337(1.391)^
b_8(Q8)	0.726(0.291)*	b_{38} (Q30)	0.271(0.23)	b_{68} (Q47 = No)	-1.329(1.441)
b_9 (Q9 = Low)	3.051(1.574)^	b_{39}(Q31 =2-3h)	1.280(0.63)*	b_{69} (Q47 = Yes)	0.045(1.462)
b_{10} (Q9 = Moderate)	-0.124(0.478)	b_{40} (Q31 = 3-4h)	1.758(1.014)^	b_{70}(Q48 = First thought)	-4.231(0.985)***
b_{11} (Q10)	0.016(0.211)	b_{41} (Q31 = <1h)	1.04(0.613)^	b_{71}(Q48 = Think through)	-2.677(0.745)***
b_{12} (Q11)	-0.033(0.252)	b_{42}(Q31 = >4h)	3.007(1.151)**	b_{72} (Q49)	0.496(0.263)^
b_{13} (Q12)	-0.148(0.239)	b_{43} (Q32 = Dining hall)	1.268(1.162)	b_{73}(I2 = Failure)	1.286(0.525)*
b_{14} (Q13)	0.056(0.240)	b_{44} (Q32 = Restaurants)	-0.052(1.18)	b_{74} (I4)	0.006(0.238)
b_{15} (Q14)	-0.070 (0.253)	b_{45} (Q32 = Home)	1.459(1.02)	<div><div><div>^ p < .01</div><div>* p < .05</div><div>** p < .01</div><div>*** p < .001</div></div><div>Median scaled residual: -0.1266</div><div>Random effects $c_i \sim N(0, 0.521^2)$</div><div>Occurrence ratio in data: 0.374</div></div>	
b_{16} (Q15)	-0.506(0.303)^	b_{46} (Q33=No)	0.153(0.604)		
b_{17}(Q16)	0.483(0.235)*	b_{47} (Q33=Some)	0.063(0.602)		
b_{18} (Q17)	0.416(0.296)	b_{48} (Q34)	-0.135(0.248)		
b_{19} (Q18)	0.408(0.255)	b_{49} (Q35)	-0.097(0.252)		
b_{20}(Q19)	-0.538(0.26)*	b_{50} (Q36)	-0.173(0.177)		
b_{21}(Q20 = Over budget)	-1.949(0.895)*	b_{51} (Q37 = 3-4)	-0.766(0.619)		
b_{22}(Q20 = Under budget)	-1.374(0.577)*	b_{52} (Q37 = >4)	-0.653(0.676)		
b_{23} (Q21 = Behind sched.)	-0.991(0.914)	b_{53} (Q37 = None)	-0.500(0.616)		
b_{24} (Q21 = On sched.)	0.015(0.835)	b_{54}(Q38)	-0.784(0.264)**		
b_{25} (Q22 = More reqs)	0.577(0.601)	b_{55} (Q39=Reqs)	0.927(0.757)		
b_{26}(Q22=reqs as planned)	1.551(0.721)*	b_{56}(Q39=Sched)	2.210(0.700)**		
b_{27} (Q23)	0.306(0.253)	b_{57} (Q40=Reqs)	1.537(1.096)		
b_{28} (Q24)	-0.03(0.274)	b_{58} (Q40=Sched)	-0.505(0.662)		
b_{29} (Q25)	-0.224(0.278)	b_{59}(Q41 = No)	-2.202(1.029)*		

Table 16: Mixed-effects logistic regression model for prediction of technical requirements failure.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	-5.444 (3.574)	b_{30} (Q26 = No)	1.117 (1.358)	b_{60} (Q41 = Yes)	0.22 (1.295)
b_1 (Q1)	0.169 (0.318)	b_{31} (Q26 = Yes)	1.821 (1.262)	b_{61} (Q42 = No)	2.313 (1.602)
b_2 (Q2)	0.081 (0.265)	b_{32}(Q27=No)	-3.054 (1.397)*	b_{62} (Q42 = Yes)	0.795 (1.64)
b_3(Q3)	0.907 (0.329)**	b_{33} (Q27=Yes)	-1.607 (1.286)	b_{63} (Q43 = Yes)	-0.944 (0.692)
b_4 (Q4)	-0.126 (0.299)	b_{34} (Q28=No)	1.967 (1.786)	b_{64} (Q44 = Yes)	0.181 (0.667)
b_5 (Q5)	-0.494 (0.323)	b_{35} (Q28=Yes)	2.167 (1.825)	b_{65} (Q45 = Yes)	0.986 (0.666)
b_6 (Q6)	0.29 (0.31)	b_{36} (Q29=No)	-0.294 (1.864)	b_{66} (Q46 = No)	1.642 (1.766)
b_7 (Q7)	-0.386 (0.3)	b_{37} (Q29=Yes)	0.361 (1.815)	b_{67} (Q46 = Yes)	-0.98 (1.79)
b_8 (Q8)	0.53 (0.355)	b_{38} (Q30)	-0.418 (0.312)	b_{68}(Q47 = No)	-3.827 (1.871)*
b_9(Q9 = Low)	3.683 (1.807)*	b_{39} (Q31 =2-3h)	1.517 (0.774)^	b_{69} (Q47 = Yes)	-3.131 (1.932)
b_{10} (Q9 = Moderate)	0.343 (0.606)	b_{40} (Q31 = 3-4h)	1.7 (1.269)	b_{70} (Q48 = First thought)	-0.155 (0.987)
b_{11} (Q10)	0.2 (0.274)	b_{41} (Q31 = <1h)	1.471 (0.818)^	b_{71} (Q48 = Think through)	-0.725 (0.848)
b_{12} (Q11)	0.229 (0.31)	b_{42} (Q31 = >4h)	-0.824 (1.397)	b_{72}(Q49)	-0.759 (0.349)*
b_{13} (Q12)	-0.264 (0.329)	b_{43} (Q32 = Dining hall)	1.542 (1.331)	b_{73}(I3 = Failure)	2.186 (0.763)**
b_{14} (Q13)	-0.249 (0.308)	b_{44} (Q32 = Restaurants)	1.476 (1.504)	b_{74} (I4)	0.019 (0.297)
b_{15} (Q14)	0.084 (0.33)	b_{45} (Q32 = Home)	0.755 (0.98)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.1108 Random effects $c_i \sim N(0, 1.128^2)$ Occurrence ratio in data: 0.254</p>	
b_{16} (Q15)	-0.667 (0.352)^	b_{46} (Q33=No)	-0.288 (0.746)		
b_{17}(Q16)	0.667 (0.333)*	b_{47} (Q33=Some)	-0.869 (0.793)		
b_{18}(Q17)	0.906 (0.37)*	b_{48} (Q34)	0.154 (0.319)		
b_{19} (Q18)	-0.17 (0.331)	b_{49} (Q35)	-0.535 (0.361)		
b_{20} (Q19)	0.099 (0.342)	b_{50} (Q36)	0.105 (0.204)		
b_{21} (Q20 = Over budget)	-0.082 (0.901)	b_{51} (Q37 = 3-4)	-1.252 (0.811)		
b_{22}(Q20 = Under budget)	1.773 (0.739)*	b_{52} (Q37 = >4)	-0.444 (0.899)		
b_{23} (Q21 = Behind sched.)	1.569 (1.38)	b_{53}(Q37 = None)	-2.405 (0.972)*		
b_{24} (Q21 = On sched.)	2.031 (1.301)	b_{54} (Q38)	-0.571 (0.351)		
b_{25} (Q22 = More reqs)	0.455 (0.766)	b_{55} (Q39=Reqs)	-0.414 (0.875)		
b_{26} (Q22=reqs as planned)	0.209 (0.855)	b_{56} (Q39=Sched)	0.489 (0.716)		
b_{27}(Q23)	0.87 (0.351)*	b_{57} (Q40=Reqs)	-3.311 (2.042)		
b_{28} (Q24)	0.16 (0.347)	b_{58} (Q40=Sched)	-0.17 (0.743)		
b_{29} (Q25)	0.646 (0.377)^	b_{59} (Q41 = No)	0.577 (1.26)		

4.2 Prediction Model Validation

To investigate the ability of the predictive models to make accurate predictions of failure outcomes, I used k -cross validation [Arlot and Celisse, 2010] with $k=10$ folds. Cross-validation is a technique to evaluate the ability of the model to generalize, that is, make accurate predictions from unknown data. To complete the validation process, I split the dataset into 10 folds. I used 9 of the folds as the training set to build the corresponding logistic regression model, and the last fold as the testing set to record the number of correct outcome predictions in that last fold. I repeated the process 10 times for each of the failure prediction models, having all folds get a chance to be the testing set, as shown in Figure 6.

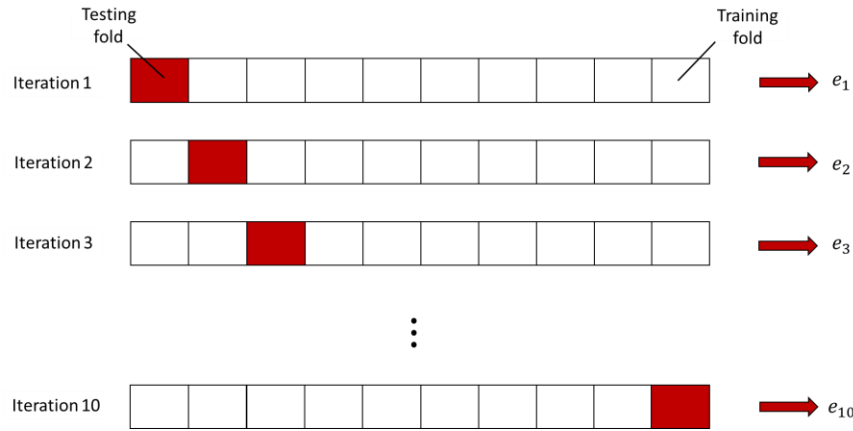


Figure 6: 10-fold cross validation process. The 267 data points (initial observations were 304, including 37 NAs) were split in 10 folds of 27 or 26. At each iteration, 240 or 241 data points were used as the training set for the logistic regression models and then the remaining 27 or 26 data points as the testing set. I recorded how many correct predictions (of the 27 or 26) the algorithm correctly identified in each iteration. I repeated the process until all observations had the chance to be included in the testing fold.

Because the true outcomes of the data points in the testing set are known, I evaluated the accuracy measure e_i in each iteration. If the model returns a predicted probability of failure greater than 50%, then I classified that as a failure. For each of the training folds, I used a confusion matrix with the predicted and actual outcomes (Table 17):

Table 17: Generic confusion matrix for logistic regression models.

	Predicted: Failure	Predicted: Not failure
Actual: Failure	n_1	n_2
Actual: Not failure	n_3	n_4

The accuracy measure is the ratio of correct outcomes identified by the model in the particular testing set, over the total outcomes:

$$e_i = \frac{n_{correct}}{n_{total}} = \frac{n_1 + n_4}{\sum_{i=1}^4 n_i} \quad (\text{Equation 11})$$

Figure 7 shows the results of the model validation process, that is, the percentage of correct predictions for each model and fold. The budget model predicted correctly, on average, $64.50 \pm 9.96\%$ of outcomes, the schedule model $60.38 \pm 13.64\%$, and the technical requirements model $66.31 \pm 10.32\%$.

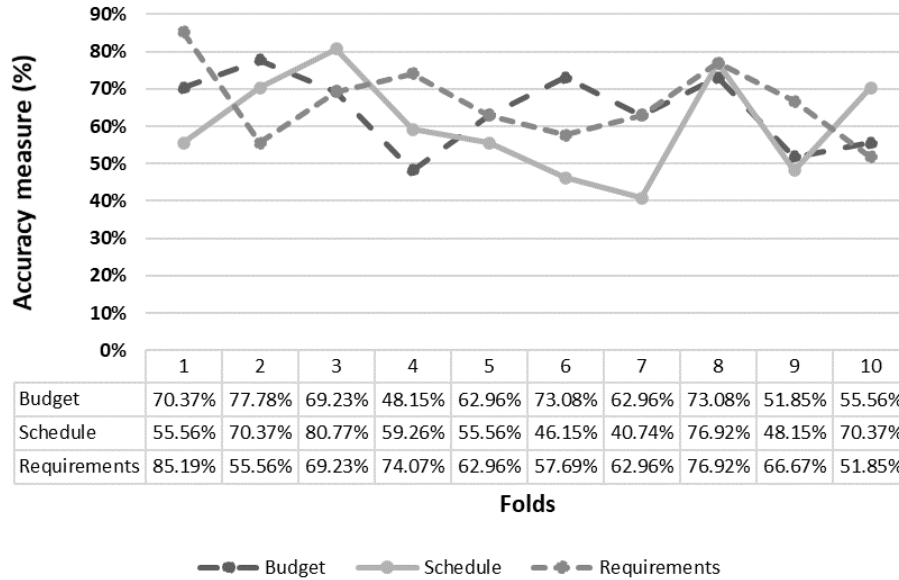


Figure 7: All three prediction models correctly predicted, on average, 60 to 65% of outcomes of unknown data. The schedule model had the highest variance of the three between the folds.

4.3 Prediction Model Reduction and Selection

In the model training so far, I considered 51 predictor variables as inputs (49 crowd signals, 1 productivity measure, and previous project state), resulting in models that show non-zero

correlations between certain inputs and specific failures, confirming that the crowd signals do have some merit in failure prediction. However, there is a drawback about the model formulation so far: the large number of inputs make the application of the predictive models not practical. If this method of predicting failures were to ever be used in an industry environment, or in a setting with larger teams, or as an app, practitioners are unlikely to expect team members to respond reliably to a 49-question-long survey every week. Therefore, as a last step in model development, I wanted to reduce the burden on the team members as much as possible, by reducing the models (i.e., arriving at a “best” model for each failure, that requires a smaller number of predictor variables). Model reduction not only reduces computational cost (because the models have fewer degrees of freedom), but also may improve how well models do at predictions (because variables that do not carry useful information for prediction are discarded).

In literature, there are a variety of methods for model reduction and selection (e.g., see [Halinski and Feldt, 1970] for a review of stepwise methods such as forward and backward selection). For these stepwise methods, a single variable is added or removed from the model at each step, according to some criterion (often R-squared or p-value). Another option is the best subsets approach [Hosmer et al. 1989]. The best subsets approach considers all possible models from all possible combinations of predictor variables and ranks them according to some criterion (i.e., if there are 10 predictor variables under consideration, best subsets will compute all possible models, that is, 2^{10} different models in this example).

In the failure prediction problem, which includes 51 predictor variables, I followed a hybrid approach: I used stepwise backwards elimination until the predictor variables were reduced to 15 (i.e., removing a maximum of 36 variables), and then a best subsets approach to arrive at the best model. The reason behind this hybrid approach was the very large number of initial variables making best subsets computationally very demanding to do from the beginning and because R includes packages that can carry out the best subsets approach with 15 variables (e.g., *bestglm*). As the criterion for my selection during both approaches, I used AIC (Akaike Information Criterion) as proposed by [Lawless and Singhal, 1987]. Table 18 shows the algorithm behind my hybrid approach for model reduction to arrive at a final best prediction model.

Table 18: Hybrid approach for model reduction and selection

$AIC_0 \leftarrow AIC(\text{initial model})$ $AIC_1 \leftarrow AIC_0$ $pred_vars \leftarrow colnames(df)$ $AIC_{history} \leftarrow AIC_0$ $Removed \leftarrow "Start"$ while $\{(AIC_1 - AIC_0 \leq 0.5) \text{ AND } (max(Removed)) \leq 36\}$ $AIC_0 \leftarrow AIC_1$ for $(i = 1, 2, \dots, length(pred_vars))$ $f \leftarrow \hat{Y}_{t+1} \sim a + bX_t(pred_vars[-i]) + c_i + \varepsilon_{it}$ $newmodel \leftarrow lme4.glmmer(f)$ $infotable[i, 1] \leftarrow AIC(newmodel)$ $infotable[i, 2] \leftarrow pred_vars[i]$ $bestmodel \leftarrow \min(infotable[, 1])$ $Removed \leftarrow c(Removed, infotable[, 2])$ $pred_vars \leftarrow pred_vars(bestmodel)$ $AIC_1 \leftarrow AIC(bestmodel)$ $AIC_{history} \leftarrow c(AIC_{history}, AIC_1)$ $bestglm(bestmodel_df, criterion = AIC, TopModels = 15)$	<p>(initial model refers to the models shown in 4.1)</p> <p>(start with all 51 predictor variables) (keep history of best AIC) (keep history of removed variables) (while loop continues with improving tolerance on AIC and a max. of 36 removed variables)</p> <p>(for loop removes one variable at a time to feed into the new model [stepwise selection]) (glmer computes the model according to f) (infotable keeps history of AIC and the prediction variable that was removed at each model)</p> <p>(bestmodel chosen based on max. AIC reduction) (Update history of AIC, removed variables, and remaining prediction variables)</p> <p>(After while loop is complete, I used bestglm to find the 15 best models with the lowest AIC)</p>
---	--

For each of the three predictive models, the remaining tables and figures in this chapter show the iterations of the stepwise removal and the results of the best subsets approach, with the final selected models.

For the budget model, the stepwise approach removed 36 predictor variables, achieving a reduction in AIC of 68, compared to the initial model (Table 19). After the best subsets approach, the best budget model includes just 10 predictor variables from the initial 51 (Table 20). Those variables correspond to questions about inability to focus on the project (*SL*, Q3), freedom on project tasks (*STND*, Q7), creativity (*CREA*, Q9), student spending estimate (*PROJS*, Q20) and confidence (*PROJSC*, Q23), student estimate of satisfying requirements (*PROJP*, Q21), previous problems resurfacing (*FSYM*, Q28), financial pressure (*FPRES*, Q38), risk perception (*RPERC*, Q39), and whether there was a failure in terms of budget the previous week (*Y_t0*, I1). The final best budget model is, on average, more accurate ($73.11 \pm 6.92\%$) and predicts correctly with less variance than the initial budget model ($64.50 \pm 9.96\%$) (Figure 9).

For the schedule model, the stepwise approach removed 36 predictor variables, achieving a reduction in AIC of 51.7, compared to the initial model (Table 22). After the best subsets approach, the best schedule model includes 15 predictor variables from the initial 51 (Table 23). The variables correspond to questions about openness to new ideas (*OPEN*, Q16), postponing or delaying obligations (*CONS*, Q17), sharing details about one's life with team members (*AGREE*, Q19), student spending estimate (*PROJS*, Q20), previous problems resurfacing (*FSYM*, Q28), number of material outputs (*OUTP*, Q30), financial pressure (*FPRES*, Q38), risk perception (*RPERC*, Q39), team members having arguments (*BANDW*, Q42), important project decisions (*FOCUS*, Q43), discussing ideas with other teams (*NOTIH*, Q45), learning new things (*CONF*, Q46), discussing unimportant matters about the project (*PARKL*, Q47), how project decisions were made (*ANCHOR*, Q48), and whether there was a failure in terms of schedule the previous week (*Y_t0*, I2). The final best schedule model is, on average, more accurate ($75.27\% \pm 9.21\%$) and predicts correctly with less variance than the initial model. ($60.38\% \pm 13.64\%$) (Figure 10).

For the technical requirements model, the stepwise approach removed 36 predictor variables achieving a reduction in AIC of 61.26 (Table 25), compared to the initial model. After the best subsets approach, the best technical requirements model includes 12 predictor variables from the initial 51 (Table 26). The variables correspond to questions about inability to focus on the project (*SL*, Q3), making meaningful progress (*IMP*, Q5), students thinking they can do progress without oversight (*AUTO*, Q8), creativity (*CREA*, Q9), feeling frustration by the team members (*NEUR*, Q15), student confidence in spending estimate (*PROJSC*, Q23), handling new problems correctly (*UNEFF*, Q26), exercising habits (*EXERC*, Q37), team members having arguments (*BANDW*, Q42), learning new things (*CONF*, Q46), student confidence in their answers (*OVERC*, Q49), and whether there was a failure in terms of technical requirements the previous week (*Y_t0*, I3). The final best technical requirements model is, on average, more accurate ($76.71 \pm 6.90\%$) and predicts correctly with less variance than the initial model ($66.31 \pm 10.32\%$) (Figure 11).

Overall, the previous status of the project was the only input variable that appeared in all three final predictive models for budget, schedule, and technical requirements, which confirms that previous project performance is a good indicator of future performance.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	
Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	
Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	Q40	
Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q49	I4	I1-3

Figure 8: Relative question importance based on their inclusion in the three final reduced models. Light grey-coded inputs appeared in 1/3 final models and dark-grey coded inputs appeared in 2/3 final models. Previous project status was the only input variable that appeared in all three final predictive models.

Table 19: Budget model stepwise predictor variable removal. The process reduced the initial AIC by 68.

<i>Iteration</i>	<i>Removed variable</i>	<i>Model AIC after removal (initial model AIC = 338.96)</i>
1	SMENG	334.57
2	EAT2	330.65
3	EXERC	326.73
4	STRN	323.09
5	BANDW	319.66
6	CONF	316.55
7	OPREF	314.05
8	COMT	312.06
9	EXTR	310.06
10	COBJ	308.07
11	TSPENT	306.08
12	COO2	304.11
13	RESO	302.15
14	AUTO	300.23
15	EXP	298.32
16	PRO	296.42
17	PROJTC	294.56
18	COMM	292.71
19	NORM	290.89
20	IMP	289.10
21	OUTP	287.62
22	UNEFF	286.06
23	MODU	284.40
24	OVERC	283.11
25	COO1	282.09
26	BUREAU	281.02
27	ANCHOR	279.71
28	PROJT	278.46
29	PROJPC	277.63
30	NTOOL	276.59
31	CONS	275.77
32	NEUR	274.62
33	NOTIH	274.00
34	PARKL	273.04
35	AMBI	271.68
36	AGREE	270.98

Table 20: Budget model best subsets results. The table shows the best 15 models by AIC, from best to worst. The best model includes 10 predictor variables with the final model having an AIC of 264.75.

SL	STND	CREA	OPEN	PROJS	PROJP	PROJSC	FSYM	EAT1	TMEET	FPRES	RPERC	FOCUS	Y_t0	PROD	AIC
In	In	In	Out	In	In	In	In	Out	Out	In	In	Out	In	Out	264.7529
In	In	In	Out	In	In	In	In	Out	Out	In	In	Out	In	In	264.8493
In	In	In	Out	In	In	In	In	Out	Out	In	In	In	In	In	265.1073
In	In	In	Out	In	In	In	In	Out	Out	In	In	In	In	Out	265.1871
In	In	In	In	In	In	In	In	Out	Out	Out	In	In	In	In	265.2358
In	In	In	In	In	In	In	In	Out	Out	In	In	Out	In	In	265.2936
In	In	In	In	In	In	In	In	Out	Out	In	In	In	In	In	265.3058
In	In	In	In	In	In	In	In	Out	Out	Out	In	Out	In	In	265.3061
In	In	In	In	In	In	In	In	Out	Out	In	In	Out	In	Out	265.403
In	In	In	Out	In	In	Out	In	Out	Out	In	In	In	In	In	265.4171
In	In	In	Out	In	In	In	In	Out	Out	Out	In	Out	In	In	265.4737
In	In	In	Out	In	In	Out	In	Out	Out	In	In	Out	In	In	265.5386
In	In	In	In	In	In	In	In	In	Out	In	In	Out	In	In	265.5576
In	In	In	In	In	In	In	In	Out	Out	In	In	In	In	Out	265.61
In	In	In	Out	In	In	In	In	In	Out	In	In	Out	In	In	265.6319

Table 21: Final budget model correlation coefficients.

Coefficient	Estimate (error)	
a	-0.711 (0.714)	
$b_1(Q3)$	0.352 (0.188)^	
$b_2(Q7)$	-0.563 (0.179)**	
$b_3(Q9=Low)$	2.090 (1.225)^	
$b_4(Q9=Moderate)$	0.869 (0.368)*	
$b_5(Q20=Over\ budget)$	-2.184 (0.867)*	
$b_6(Q20 = Under\ budget)$	0.133 (0.403)	
$b_7(Q21 =Fewer\ reqs.)$	-1.746 (0.63)**	
$b_8(Q21 = More\ reqs.)$	0.025 (0.497)	
$b_9(Q23)$	0.335 (0.194)	
$b_{10}(Q28 = No)$	-1.452 (0.639)*	
$b_{11}(Q28=Yes)$	-0.982 (0.675)	
$b_{12}(Q38)$	-0.343 (0.185)^	
$b_{13}(Q39=Reqs)$	-0.449 (0.475)	
$b_{14}(Q39 =Sched.)$	0.845 (0.463)^	
$b_{15}(I1=Failure)$	1.261 (0.369)***	

$^{\wedge} p < .01$
 $* p < .05$
 $** p < .01$
 $*** p < .001$

Median scaled residual: -0.273
Random effects $c_i \sim N(0, 0.541^2)$
Occurrence ratio in data : 0.254

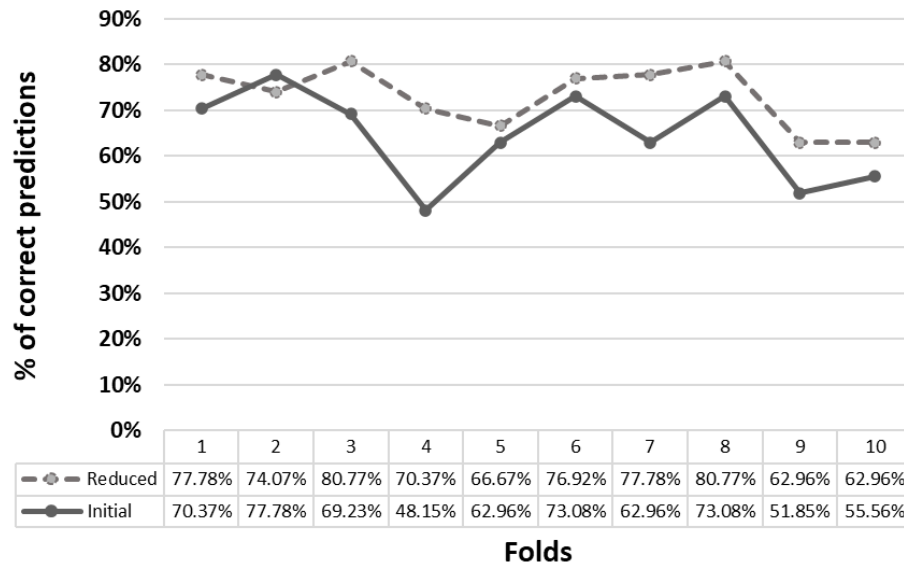


Figure 9: The final best budget model is, on average, more accurate ($73.11 \pm 6.92\%$) and predicts correctly with less variance than the initial budget model ($64.50 \pm 9.96\%$).

Table 22: Schedule model stepwise predictor variable removal. The process reduced the initial AIC by 51.7.

<i>Iteration</i>	<i>Removed variable</i>	<i>Model AIC after removal (initial model AIC = 371.20)</i>
1	EXERC	367.03
2	EAT2	363.20
3	STRM	360.27
4	STND	358.27
5	PROJTC	356.27
6	PROD	354.28
7	MODU	352.28
8	COBJ	350.30
9	EXP	348.38
10	COMM	346.47
11	RESO	344.56
12	PRO	342.66
13	TMEET	340.85
14	PROJPC	339.09
15	UNEFF	337.53
16	IMP	336.06
17	COMT	334.63
18	TSPENT	333.39
19	COO2	332.10
20	SMENG	330.84
21	EAT1	328.37
22	BUREAU	326.24
23	OPREF	324.65
24	EXTR	324.17
25	COO1	323.62
26	PROJP	322.83
27	NEUR	322.03
28	PROJSC	321.33
29	NORM	320.81
30	NTOOL	320.44
31	OVERC	320.53*
32	AMBI	320.71*
33	PROJT	319.66
34	AUTO	319.68
35	SL	319.74
36	CREA	319.50

*Even if AIC slightly increased in these iterations, it is within acceptable tolerance for the algorithm to continue searching further to arrive at models with lower AIC.

Table 23: Schedule model best subsets results. The table shows the best 15 models by AIC, from best to worst. The best model includes 15 predictor variables with the final model having an AIC of 315.5.

OPEN	CONS	AGREE	PROJS	FSYM	OUTP	FPRES	RPERC	BANDW	FOCUS	NOTIH	CONF	PARKL	ANCHOR	Y_t0	AIC
In	In	In	In	In	In	In	In	In	In	In	In	In	In	In	315.4983
In	In	In	In	In	Out	In	In	In	In	In	In	In	In	In	316.2376
In	In	In	In	In	In	In	In	In	In	In	In	Out	In	In	316.6819
In	Out	In	In	In	In	In	In	In	In	In	In	In	In	In	317.3208
In	In	In	In	In	Out	In	In	In	In	In	In	Out	In	In	317.5584
In	Out	In	In	In	Out	In	In	In	In	In	In	In	In	In	318.0073
In	Out	In	In	In	In	In	In	In	In	In	In	Out	In	In	318.0829
Out	In	In	In	In	Out	In	In	In	In	In	In	In	In	In	318.6929
In	In	In	In	Out	Out	In	In	In	In	In	In	In	In	In	318.7417
In	In	In	Out	In	In	In	In	In	In	In	In	Out	In	In	318.7477
Out	In	In	In	In	In	In	In	In	In	In	In	In	In	In	318.7962
In	In	In	In	Out	In	In	In	In	In	In	In	In	In	In	318.7989
Out	In	In	In	In	In	In	In	In	In	In	In	Out	In	In	318.8702
Out	In	In	In	In	Out	In	In	In	In	In	In	Out	In	In	318.8939
In	Out	In	In	In	Out	In	In	In	In	In	In	Out	In	In	318.9099

Table 24: Final schedule model correlation coefficients.

Coefficient	Estimate (error)	
a	0.773 (1.509)	
$b_1(Q16)$	0.379 (0.164)*	
$b_2(Q17)$	0.341 (0.177)^	
$b_3(Q19)$	-0.413 (0.172)*	
$b_4(Q20=Over\ budget)$	-1.291 (0.629)*	
$b_5(Q20 = Under\ budget)$	-0.791 (0.388)*	
$b_6(Q28 = No)$	0.494 (0.661)	
$b_7(Q28=Yes)$	1.346 (0.685)*	$\wedge p < .01$
$b_8(Q30)$	0.271 (0.165)	$* p < .05$
$b_9(Q38)$	-0.464 (0.176)**	$** p < .01$
$b_{10}(Q39=Reqs)$	0.677 (0.451)	$*** p < .001$
$b_{11}(Q39 =Sched.)$	1.265 (0.448)**	
$b_{12}(Q42 = No)$	-1.053 (0.735)	Median scaled residual: -0.279
$b_{13}(Q42=Yes)$	-1.915 (0.802)*	Random effects $c_i \sim N(0, 0.316^2)$
$b_{14}(Q43=Yes)$	-1.025 (0.372)**	Occurrence ratio in data : 0.374
$b_{15}(Q45=Yes)$	0.838 (0.362)*	
$b_{16}(Q46 = No)$	2.057 (1.062)^	
$b_{17}(Q46=Yes)$	0.976 (1.029)	
$b_{18}(Q47 = No)$	-1.707 (1.071)	
$b_{19}(Q47=Yes)$	-1.027 (1.066)	
$b_{20}(Q48=First\ thought)$	-2.387 (0.620)***	
$b_{21}(Q49=Think\ through)$	-1.821 (0.478)***	
$b_{22}(I2=Failure)$	1.069 (0.333)**	

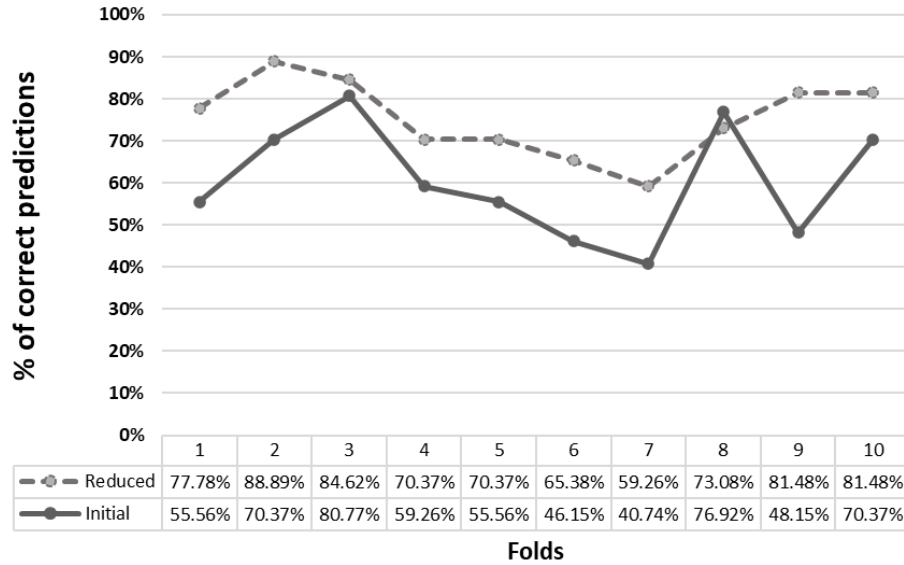


Figure 10: The final best schedule model is, on average, more accurate ($75.27\% \pm 9.21\%$) and predicts correctly with less variance than the initial model ($60.38\% \pm 13.64\%$).

Table 25: Technical requirements model stepwise predictor variable removal. The process reduced the initial AIC by 61.26.

<i>Iteration</i>	<i>Removed variable</i>	<i>Model AIC after removal (initial model AIC = 337.82)</i>
1	EAT1	333.96
2	EAT2	330.24
3	PROJP	326.60
4	AMBI	323.56
5	BUREAU	320.83
6	FSYM	318.27
7	ANCHOR	316.21
8	COO1	314.21
9	NORM	312.22
10	TSPENT	310.26
11	EXP	308.38
12	COMM	306.51
13	PROJTC	304.65
14	NTOOL	302.79
15	PROD	300.92
16	MODU	299.13
17	AGREE	297.42
18	COBJ	295.74
19	RESO	294.14
20	PRO	292.66
21	FOCUS	291.55
22	NOTIH	290.15
23	SMENG	288.97
24	PROJT	287.42
25	FPRES	285.99
26	STRM	284.62
27	STND	282.98
28	EXTR	281.73
29	COMT	280.83
30	RPERC	280.21
31	COO2	279.24
32	OUTP	278.48
33	OPREF	278.69*
34	PARKL	277.98
35	CONS	277.24
36	OPEN	276.56

Table 27: Final technical requirements model correlation coefficients.

Coefficient	Estimate (error)	
a	-3.755 (1.531)*	
b₁(Q3)	0.697 (0.212)**	
b₂(Q5)	-0.405 (0.197)*	
b₃(Q8)	0.472 (0.204)*	
b₄(Q9=Low)	2.672 (1.168)*	
b₅(Q9=Moderate)	0.43 (0.39)	
b₆(Q15)	-0.284 (0.212)	
b₇(Q23)	0.651 (0.22)**	
b₈(Q26=No)	1.419 (0.856)^	
b₉(Q26=Yes)	2.336 (0.785)**	
b₁₀(Q37=3-4)	-0.85 (0.465)^	
b₁₁(Q37=>4)	-0.603 (0.523)	
b₁₂(Q37=None)	-1.726 (0.55)**	
b₁₃(Q42=No)	0.86 (0.855)	
b₁₄(Q42=Yes)	-0.134 (0.937)	
b₁₅(Q46=No)	0.574 (1.255)	
b₁₆(Q46=Yes)	-0.674 (1.274)	
b₁₇(Q49)	-0.458 (0.195)*	
b₁₈(I3=Failure)	1.479 (0.408)***	

^ p < .01
 * p < .05
 ** p < .01
 *** p < .001

Median scaled residual: -0.247
 Random effects $c_i \sim N(0, 0.811^2)$
 Occurrence ratio in data: 0.254

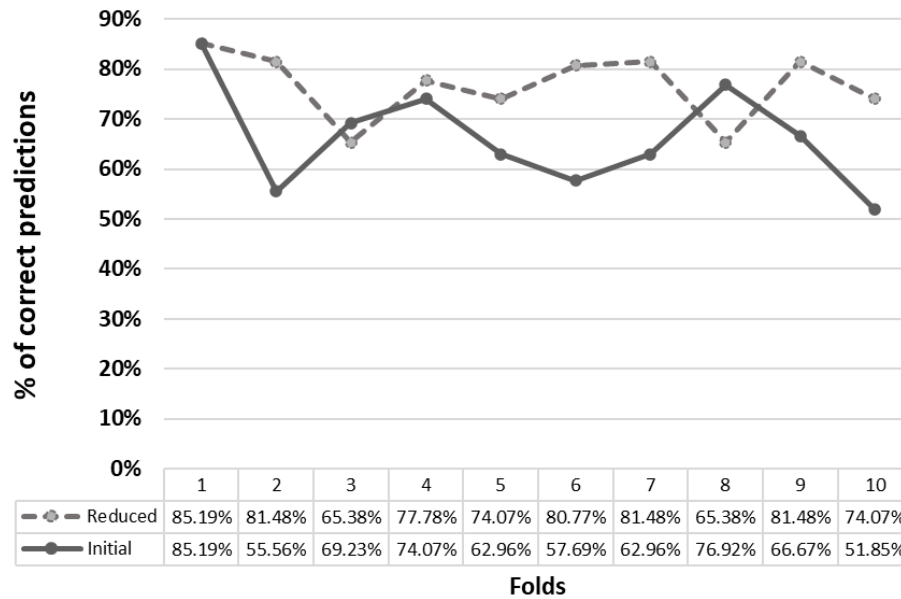


Figure 11: The final best technical requirements model is, on average, more accurate ($76.71 \pm 6.90\%$) and predicts correctly with less variance than the initial model ($66.31 \pm 10.32\%$).

5. EXPERIMENT II: TARGETED FEEDBACK TO PREVENT FAILURES IN SYSTEMS ENGINEERING

This chapter describes the second experiment to provide targeted feedback to the student teams and evaluate whether the feedback was helpful at reducing failure occurrences in student projects. The experimental procedures were approved as an exempt study by Purdue's Institutional Review Board (IRB) with protocol #2020-1393 and title “Targeted Feedback to Prevent Systems Engineering Failures”. I collected the data for the second experiment during the Spring semester of 2021. Chapter 5 is organized as follows: Section 5.1 provides a description of the experimental setup and design. Section 5.2 discusses the feedback process. Section 5.3 shows a derivation for the overall probability of failure for a project team, which is part of the feedback. Section 5.4 focuses on the development of the feedback statements and their associated rules. Section 0 concludes with the evaluation of the feedback using statistical testing and qualitative survey metrics.

5.1 Experimental Setup and Design

For the second experiment, I followed a similar recruiting process as for experiment I: I asked the students of engineering design courses to volunteer as respondents to a brief survey at the end of each week, answering a set of questions (student crowd signals). I provided a weekly \$20 gift card incentive for one randomly selected student each week. There were two additions to the student survey, compared to the one from experiment I. The additions were the weekly feedback statements and three additional questions for the students to provide their evaluation of the feedback. At the same time, the instructors of each course responded to a separate survey at the end of each week since I needed their assessment of the project status to use my predictive models. The criterion for student recruitment in the study was to be enrolled in an engineering course that includes a team design project and to be above 18 years of age. The criterion for instructor recruitment was to monitor student teams closely, to be able to accurately provide the progress of each project.

Due to the COVID-19 restrictions at the time of the experiment, I handled all recruiting processes and data collection entirely online. To motivate the students and to explain the research in an approachable way, I created a recruiting video and flyer. The video was 2 minutes and 50 seconds long, and introduced me, as well as the process the students should follow. The flyer (Figure 12) was similar to the one used for Experiment I, and included contact details as well as the QR code for students to access the survey with their mobile devices. In communication with the instructors, I was also able to send some email reminders to encourage participation of the students. For confidentiality purposes, I followed the same approach with student usernames as in Experiment I (anonymous link for the survey, and no identifiable information collected from the students).

Targeted Feedback to Prevent Systems Engineering Failures

Have You Ever Wondered Why
Your Team Project Did Not Turn
Out So Well?


Or what you can do to improve
your team's performance?

We did too... and we can try to help you. We have
generated a series of helpful feedback for you, that we
think will improve your experience in team-based
engineering projects.

You can be a part of this!
We ask you to respond to our survey questions about
various aspects of the student project you are part of.
Every week, you will get helpful suggestions as
feedback, to improve your (and your team's)
performance!

Are my responses confidential?
Absolutely! When you participate, pick a username you
can remember, and use it for the entire semester.

**OK, take me to the
survey**


**(QR code and web
link to survey)**

**ANSWER SURVEY
QUESTIONS ABOUT
YOUR PROJECT —
RECEIVE HELPFUL
FEEDBACK**

**TIME COMMITMENT:
10 MINS PER WEEK**

**\$20 GIFT CARD
DRAWING EVERY WEEK**

**HAVE QUESTIONS?
CONTACT US!**

Karen Marais
AAE Professor
kmarais@purdue.edu

Georgios Georgalis
PhD Candidate
ggeorgal@purdue.edu

Figure 12: Recruitment flyer as distributed to the students during the recruitment process of Experiment II. 10 random students won a \$20 gift card.

The experiment included two treatment groups: the student teams that received targeted feedback statements and teams that received non-targeted feedback statements. I used two treatment groups because I wanted to isolate the effect of the targeted feedback statements on failure occurrences, and I had to use a valid comparison. Comparing to teams that received no feedback would not make a valid comparison because any statistical significance could be because of the effect of

feedback statements in general, rather than the targeted feedback statements I developed. Comparing failure rates to teams that received non-targeted feedback, however, would isolate the effect of my targeted feedback process against feedback statements that do not necessarily address the failure causes the team is more prone to. The feedback statements were all positive, encouraging, and promoted good team practices for both treatment groups.

The teams that received targeted feedback got three statements from those that corresponded to specific rules that I explain in detail in section 5.4. The rules help distinguish between feedback statements that address the failure causes the team is prone to (“targeted feedback”) and those statements that do not necessarily do so (“non-targeted feedback”). The teams that received the non-targeted feedback got three feedback statements from the ones that did not apply to them based on the rules. Both treatment groups received the truthful predicted probabilities of failure that the predictive models output based on their responses.

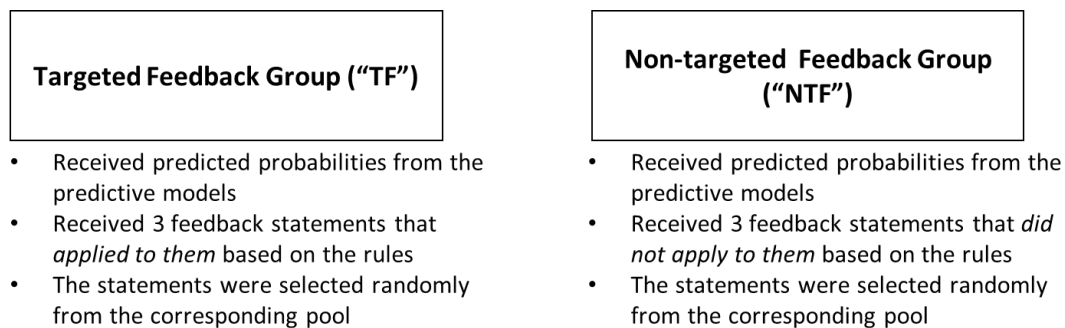


Figure 13: The two treatment groups used in Experiment II. The difference was the process for the feedback statements: one group received from statements that applied to them based on the rules, whereas the other group received from statements that did not apply to them based on the rules.

In total, I collected data from 14 different design project teams. The student teams were enrolled in three different courses at Purdue University. All data collection occurred during the same 9-week period. For these courses, the student teams worked on aircraft design, propulsion design-build-test, or spacecraft design. In total, 53 students participated at least one time in the survey. Table 28 shows a summary of the data collected per course.

Table 28: Summary of data collection for experiment II during the Spring '21 semester. Student teams typically included 4-6 team members. The projects included both hardware and software deliverables as well as progress and final reports.

<i>Course</i>	<i># projects</i>	<i># student observations (excl. NAs)</i>	<i># instructor observations</i>
Aerospace #1	8	69	72 (8 projects for 9 weeks)
Aerospace #2	2	26	4 (2 projects for 2 weeks)
Aerospace #3	4	37	36 (4 projects for 9 weeks)

5.2 Feedback process

The feedback included two parts: the first part to alert the project team of upcoming failures and the second part to provide the feedback statements.

For the alert part, I provided the predicted failure probability from the reduced models discussed earlier in Section 4.3. Given an input set of crowd signals from the students and the state of the project from the instructor during a given week, the models can predict the likelihood of the team to have a budget, schedule, or technical requirements failure the following week.

For the statement part, I provided recommendations based on the treatment group each project team belonged to, as described earlier.

I developed the feedback statements based on the idea discussed in the introduction (Figure 2): to use the correlations between failure causes and crowd signals to generate feedback statements that attempt to improve student behavior to address a particular failure cause. In Section 3.2, I discussed the process of finding such correlations for the failure cause FC1: *Failed to consider a design aspect*. I used the exact same process for the remaining failure causes and created a matrix (the “Crowd Signal—Failure Cause” or “CS—FC” matrix) to guide the feedback statement generation process. For every *crowd signal* + *failure cause* pair that had an existing correlation, I created an associated feedback statement based on the type of correlation.

Each statement came with a rule that distinguishes whether the statement is applicable (“targeted”) to a particular student team (i.e., addresses an area they may be weak at, based on their responses). The rules were in place to remove any potential bias caused by my ability to make judgments as to which statement is a good fit for each team, due to my knowledge of how student teams work. The rules were expressed in relation to the associated crowd signal. The rules also enable the process for the two treatment groups. The “TF” group gets statements when their responses satisfy the rules (i.e., statements that address the failure causes the team is more prone to), whereas the “NTF” group gets statements from those that their responses do not satisfy the rules (i.e., statements that are positive and encouraging, but do not necessarily address the failure causes the team is more prone to).

The feedback, in general, was different for every team and every week, because the inputs were different for the predictive models (and therefore the predicted probabilities and applicable rules that were satisfied). The only reason the feedback would remain the same, is if the students of a particular team did not provide new responses for that week (so there was no updated predictions or feedback). To generate the first feedback statement, one week’s data collection was required in the beginning. The following steps summarize the process:

1. *Week 1*: Collected data from students/instructors.
2. *Week 2*: Students saw 1st feedback message; collected new data from students/instructors.
3. *Week 3*: Students saw new feedback message and evaluated previous feedback message; collected new data from students/instructors.
4. Repeat until *Week 9* and record final project performance from instructors.

To ensure each team sees the correct feedback message that applies to them, I used a Qualtrics feature that allowed me to display particular messages conditional on which team the student selects when they respond to the survey. The feedback message had the following format for both treatment groups:

“Based on models we built with data from previous teams that received no feedback and the responses from your team members from last week: We predict that you have $[\hat{P}_{i,t+1}^{(k)}\%]$ chance of having a failure in terms of [metric i]. To improve your team’s chances of success, we suggest $[R_{Q-FC}^{(k)}]$ ”.

The measures in brackets were edited for the different teams during the duration of the experiment. The index i of the probability $\hat{P}_{i,t+1}^{(k)}$ reflects one of the success metrics: $i = 1$ corresponds to a budget failure, $i = 2$ to a schedule failure, and $i = 3$ to a technical requirements failure. The superscript reflects the team k the prediction corresponded to. $R_{Q-FC}^{(k)}$ was the set of feedback statements that applied to team k . Figure 14 shows the steps to generate the necessary measures to provide feedback to the project teams.

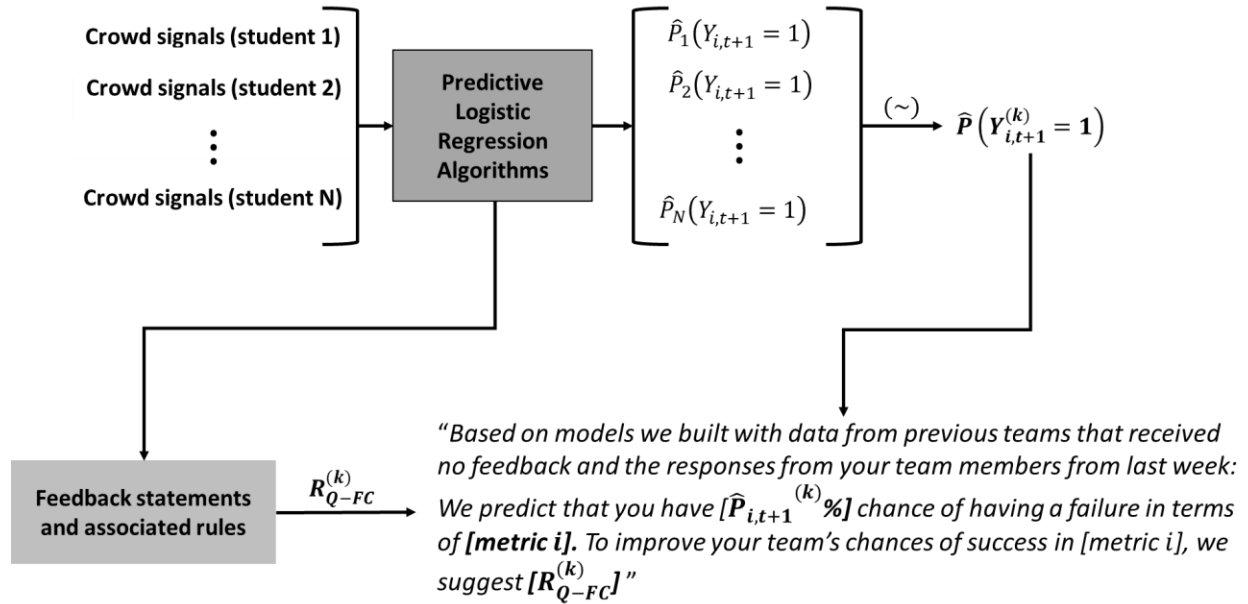


Figure 14: The dynamic feedback process. At each week t , through the failure predictive models, I calculated the probability of failure for each metric and team for the following week $t+1$. Using this probability and the associated feedback rules, I provided recommendations to the team members.

5.3 Calculation of Overall Probability of Failure for a Project Team

Figure 14 shows a process of calculating the probability of failure for a metric for the entire team, given the failure probabilities calculated from the individual student responses (\sim). The way I built the models, responses from different students who are in the same team will output different predicted probabilities of failure for the same project. However, the goal of the alert part of the feedback is to provide one overall probability of failure for the entire team. I show here the derivation of the overall probability of failure for the entire project, given the individual and independent responses of students that are part of the team.

If I had built a model based on grouped student responses by team using some arbitrary grouping scheme, that model would output the predicted probability of failure in terms of metric j given the grouped responses from the team:

$$\begin{aligned} \hat{P}(Y_{j,t+1} = 1 | Q_1^{(1...N)} = g_1, Q_2^{(1...N)} = g_2, \dots, Q_n^{(1...N)} = g_n) \\ = \hat{P}(Y_{j,t+1} = 1 | Q_{1...n}^{(1...N)} = g_{1...n}) \end{aligned} \quad (\text{Equation 12})$$

Where $Y_{j,t+1}$ corresponds to a failure in terms of metric j for week $t + 1$. $Q_1^{(1...N)}$ represents the student responses to the first question that are grouped according to some scheme $g(\cdot)$ with value g_1 , $Q_2^{(1...N)}$ represents the grouped student responses to the second question with value g_2 , etc. for all N students. I use $Q_{1...n}^{(1...N)} = g_{1...n}$ as a shortened notation for the grouped responses to all questions.

From Bayes' theorem and the law of total probability (by conditioning the probability that forms in the denominator using the two possible failure outcomes):

$$\begin{aligned} & \hat{P}(Y_{j,t+1} = 1 | Q_{1...n}^{(1...N)} = g_{1...n}) \\ &= \frac{\hat{P}(Y_{j,t+1} = 1, Q_{1...n}^{(1...N)} = g_{1...n})}{\hat{P}(Y_{j,t+1} = 1, Q_{1...n}^{(1...N)} = g_{1...n}) + \hat{P}(Y_{j,t+1} = 0, Q_{1...n}^{(1...N)} = g_{1...n})} \end{aligned} \quad (\text{Equation 13})$$

Since the grouped responses come from the individual student responses and the probability of overall project failure comes from the failure per each individual student, the numerator of the RHS of equation (13) becomes:

$$\hat{P}(Y_{j,t+1} = 1, Q_{1...n}^{(1...N)} = g_{1...n}) = \hat{P} \left(\begin{matrix} \{Y_{j,t+1}^{(1)} = 1\} & \{q_{1...n}^{(1)} = Q_{1...n}^{(1)}\} \\ \{Y_{j,t+1}^{(2)} = 1\} & \{q_{1...n}^{(2)} = Q_{1...n}^{(2)}\} \\ \vdots & \vdots \\ \{Y_{j,t+1}^{(N)} = 1\} & \{q_{1...n}^{(N)} = Q_{1...n}^{(N)}\} \end{matrix} \right) \quad (\text{Equation 14})$$

Where $q_{1...n}^{(1)}$ represents the answers of the first student to questions $Q_{1...n}$, $q_{1...n}^{(2)}$ represents the answers of the second student to questions $Q_{1...n}$ etc.

Assuming that the responses and failure probability of each student is independent of the other students (which is reasonable because the questions primarily capture the individual student's opinions, behaviors, and actions), the RHS of equation (14) becomes:

$$\hat{P} \left(\begin{matrix} \{Y_{j,t+1}^{(1)} = 1\} & \{q_{1...n}^{(1)} = Q_{1...n}^{(1)}\} \\ \{Y_{j,t+1}^{(2)} = 1\} & \{q_{1...n}^{(2)} = Q_{1...n}^{(2)}\} \\ \vdots & \vdots \\ \{Y_{j,t+1}^{(N)} = 1\} & \{q_{1...n}^{(N)} = Q_{1...n}^{(N)}\} \end{matrix} \right) = \prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 1, q_{1...n}^{(i)}) \quad (\text{Equation 15})$$

Using Bayes' theorem for the individual student probabilities inside the product:

$$\prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 1, q_{1...n}^{(i)}) = \prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 1 | q_{1...n}^{(i)}) P(q_{1...n}^{(i)}) \quad (\text{Equation 16})$$

And equivalently for the scenario a failure did not occur ($Y_{j,t+1} = 0$):

$$\begin{aligned} \hat{P}(Y_{j,t+1} = 0, Q_{1...n}^{(1...N)} = g_{1...n}) &= \prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 0, q_{1...n}^{(i)}) \\ &= \prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 0 | q_{1...n}^{(i)}) P(q_{1...n}^{(i)}) \end{aligned} \quad (\text{Equation 17})$$

Back-substituting equations 16 and 17 in equation 13 gives:

$$\begin{aligned} &\hat{P}(Y_{j,t+1} = 1 | Q_{1...n}^{(1...N)} = g_{1...n}) \\ &= \frac{\prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 1 | q_{1...n}^{(i)})}{\prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 1 | q_{1...n}^{(i)}) + \prod_{i=1}^{i=N} \hat{P}(Y_{j,t+1}^{(i)} = 0 | q_{1...n}^{(i)})} \end{aligned} \quad (\text{Equation 18})$$

The quantities on the RHS of equation 18 can be calculated from the regression models, since for each student the model gives the probability of failure given their individual responses $q_{1...n}^{(i)}$ for all

questions and for all N students that are in the same team. The complement case, when $Y_{j,t+1} = 0$, can also be calculated from the regression models by using the law of total probability since there are only two outcomes in each prediction: failure or no failure.

The formulation in equation 18 represents the overall probability of failure in terms of any of the three metrics for a project, given individual responses from various team members each week.

5.4 Feedback Statement Development and Rules

To develop the feedback statements in a structured and consistent manner, I used the correlations between failure causes and crowd signals to guide the process. The goal was for the targeted feedback statements to recommend actions and behaviors that may improve the underlying failure causes a team is most prone to, and therefore reduce the occurrences of project failures (for the “TF” group). To find these correlations between failure causes and crowd signals, I used the same modeling approach as described for FC1: *Failed to consider a design aspect* in Section 3.2. The detailed coefficients values for each model can be found in Appendix A.

With 10 failure causes and 49 crowd signals, to best facilitate the feedback statement generation process, I summarized the information I needed in a matrix form, which I named the “Crowd Signal—Failure Cause” or “CS—FC” matrix (Figure 15). The rows of the matrix are the 49 crowd signals questions, the columns are the failure causes FC1 to FC10, and each cell shows the type of correlation (“+” indicates a positive correlation, “−” indicates a negative correlation, and “0” indicates no correlation). Positive correlation means the corresponding model included a positive coefficient with a p-value of 0.05 or less, negative correlation means the model included a negative coefficient with a p-value of 0.05 or less, and no correlation means the p-value for the corresponding coefficient was more than 0.05. FC3: *Failed to form a contingency plan* (occurrence ratio = 0.107) and FC9: *Violated procedures* do not have a model (occurrence ratio = 0.064) because they did not occur enough times for the models to converge and therefore, I excluded them from the feedback generation process.

	Failure causes									
	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10
Q1	0	0	~	0	0	-	+	0	~	0
Q2	0	+	~	0	0	+	0	0	~	0
Q3	0	0	~	0	0	0	+	+	~	0
Q4	-	0	~	0	0	0	-	0	~	0
Q5	0	0	~	0	0	+	0	0	~	0
Q6	-	0	~	0	0	0	0	0	~	0
Q7	0	0	~	0	0	0	0	0	~	0
Q8	0	0	~	0	0	0	0	0	~	0
Q9	0	0	~	0	0	(Mod) -	0	0	~	0
Q10	0	0	~	0	+	0	0	0	~	+
Q11	0	0	~	0	-	0	0	0	~	0
Q12	0	+	~	0	0	0	0	0	~	0
Q13	0	0	~	0	0	0	0	-	~	0
Q14	0	0	~	0	0	0	0	0	~	0
Q15	0	0	~	0	-	0	0	0	~	0
Q16	0	0	~	0	0	-	+	0	~	0
Q17	0	0	~	0	0	+	0	0	~	0
Q18	0	0	~	0	0	0	0	0	~	0
Q19	0	0	~	0	0	0	-	0	~	0
Q20	0	0	~	0	(Over) -	0	0	0	~	0
Q21	0	(Behind, On) -	~	0	0	0	(Behind) -	(Behind, On) -	~	0
Q22	0	0	~	0	(As plan) -	0	0	0	~	0
Q23	0	0	~	0	+	0	0	0	~	0
Q24	0	0	~	+	+	0	0	+	~	0
Q25	0	0	~	0	0	0	+	0	~	0
Q26	(No) -	0	~	0	0	(No) -	0	0	~	0
Q27	0	0	~	0	0	0	0	(Yes) +	~	0
Q28	(No) -	0	~	0	0	0	0	0	~	0
Q29	(No) +	0	~	0	0	0	0	0	~	0
Q30	0	0	~	0	0	0	+	0	~	0
Q31	0	(<1h) -	~	0	(3-4h) -	0	(2-3h) -	0	~	(2-3h) -
Q32	0	0	~	0	(Home) -	(Hall) -	0	0	~	(Home) +
Q33	0	(Some) +	~	0	(Some) +	0	(No, Some) +	0	~	(Some) +
Q34	0	0	~	0	-	0	0	0	~	0
Q35	0	+	~	+	0	+	0	+	~	0
Q36	-	0	~	0	+	0	0	0	~	0
Q37	0	0	~	0	0	0	0	(>4) -	~	0
Q38	0	0	~	0	0	0	0	0	~	0
Q39	0	0	~	0	(Reqs, sched) -	0	0	(Reqs) -	~	0
Q40	(Reqs) -	(Reqs) +	~	0	0	0	(Reqs, sched) -	0	~	0
Q41	0	(Yes) +	~	0	+	0	0	(Yes) +	~	0
Q42	(No) -	0	~	0	0	(No) -	0	(No) -	~	0
Q43	0	0	~	0	0	0	0	0	~	0
Q44	0	0	~	0	0	0	0	0	~	0
Q45	0	(Yes) -	~	0	0	0	0	0	~	0
Q46	0	0	~	(No) -	0	0	0	0	~	0
Q47	0	(No) +	~	0	0	(Yes) -	0	0	~	0
Q48	0	0	~	0	(Think) +	0	0	0	~	(Think) +
Q49	0	0	~	0	+	0	0	+	~	0

Figure 15: The crowd signal–failure cause correlation matrix. “+” indicates a positive correlation, “-” indicates a negative correlation, and “0” indicates no correlation. When the questions have categorical answers, the correlation is labeled with the answer that it corresponds to. FC3 and FC9 are excluded due to low occurrence ratios in the corresponding model training data sets.

The process of developing the feedback statements is to create a statement for each question/failure cause cell that includes a non-zero value in the “CS—FC” matrix.

To illustrate the development process, I provide here one example for the feedback statements that I created from the second row of the CS—FC matrix (Q2). Q2 measures proactivity by asking “During the past week, how many times did you attempt to get involved with a project-related task that was outside your immediate responsibility?”. Q2 has a positive correlation with FC2: *Used inadequate justification* and FC6: *Inadequately communicated*. The interpretation of the positive correlation means that the more times students attempt to get involved with tasks outside their responsibilities, we should expect more occurrences of poor communication or inadequate justification of a decision.

The recommendation in this case, starts by putting the student in the situation addressed by Q2, and then makes a suggestion to avoid FC2 and FC6: *“When you offer to get involved or help with a task that another team member works on: make sure that you communicate well about the level of your involvement, specify exactly what is expected of you, and let them know of your thought process for any decisions you make.”*

The rule associated with this recommendation was “Provide recommendation if average team response to Q2 is more than 2 times”. Q2 is an integer and the reasoning behind this rule was to try and help teams where their members become involved outside their tasks, making them vulnerable to FC2 and FC6. The remaining rules followed a similar thought process: if the correlation was positive (i.e., as the crowd signal increases, the failure cause likelihood increases), then the rule is given once the crowd signal exceeds a value. If the correlation was negative (i.e., as the crowd signal increases, the failure cause likelihood decreases), then the rule was given once the crowd signal drops below a value. The rule threshold values reflect the median possible response where possible (e.g., for Likert-scale answers that value is 2.5). For the categorical answers, the values reflect a majority response and depend on which answer showed the correlation (e.g., if there is a negative correlation associated with “Moderate and low creativity” in Q9, then the threshold value is “High creativity”).

I developed all the feedback statements in a similar manner (Table 29). In some cases, even if there is a correlation, a feedback statement is not possible, either because of the question itself or because the student does not have control over that measure in the project timeframe. For example, Q1 measures how many previous projects the student has been a part of, which is not possible for them to change during the progress of a project, and so Q1 did not generate any feedback statements (marked as N/A).

Table 29: The 35 feedback statements and associated rules for each existing correlation in the crowd signal–failure cause correlation matrix.

Correlation Matrix Cell		Feedback Statement	Rule
Q1–FC6/Q1–FC7		N/A	N/A
Q2–FC2/Q2–FC6	R1	When you offer to get involved or help with a task that another team member works on: make sure that you communicate well about the level of your involvement, specify exactly what is expected of you, and let them know of your thought process for any decisions you make.	Average team response ≥ 2
Q3–FC7/Q3–FC8	R2	When you find yourself unable to focus and you have upcoming testing or important updates that are related to the safety of the project, ask your teammates for assistance to minimize mistakes.	Average team response ≥ 2.5
Q4–FC1/Q4–FC7	R3	When you are in the design or testing phase of a component or part, discuss with your teammates to confirm you have thought of all potential features or requirements that are crucial, especially if the component is to be integrated with other systems.	Average team response ≤ 2.5
Q5–FC6	R4	When you made a lot of progress in a week, make sure everyone is aware about the accomplishments and the completed work.	Average team response ≥ 2.5
Q6–FC1	R5	Try to have some discussion about what your teammates worked and achieved every week, as this will help you understand the whole system better and reduce the chance of missing a key design aspect.	Average team response $\leq 50\%$
Q7		N/A	N/A
Q8		N/A	N/A
Q9–FC6	R6	If you find your team proposing a lot of ideas about a project design, make an attempt to communicate well and think through these options with each other instead of simply listing many of them.	Majority response = High creativity
Q10–FC5/Q10–FC10	R7	When you have to do mostly independent work during the week, keep good documentation about what you are doing so it is clear to the entire team how all the resulting work is to be integrated.	Average team response $\geq 50\%$
Q11–FC5	R8	In the beginning of the week, have a discussion with all team members present and clarify exactly what the objectives are for the week.	Average team response ≤ 2.5
Q12–FC2	R9	If you find that you are not working well with each other, start by discussing about the reasoning behind all your technical work, which may allow you to jumpstart your teamwork.	Average team response ≥ 2.5

Table 29 continued

<i>Correlation Matrix Cell</i>		<i>Feedback Statement</i>	<i>Rule</i>
Q13–FC8	R10	Ask your instructor for any necessary resources you need to ensure the safety of the project, such as to come up with a mitigation measure or a redundant component.	Average team response ≤ 2.5
Q14		N/A	N/A
Q15–FC5	R11	When there is lack of arguments in the team, make sure you are not becoming complacent and put effort into properly recording everything you do as a team.	Average team response ≤ 2.5
Q16–FC6	R12	Try to spend some time every week in idea-generation sessions where everyone proposes a new idea about how to improve your project design.	Average team response ≤ 2.5
Q16–FC7	R13	While discussing new ideas or decisions for your project with your team, it is important that you are spending enough time and effort on project-critical activities such as testing.	Average team response ≥ 2.5
Q17–FC6	R14	If you have many obligations in a given week, communicate those with your team so everyone knows when you will not be available or too busy.	Average team response ≥ 2.5
Q18		N/A	N/A
Q19–FC7	R15	Make an effort to get to know your teammates, especially those that you frequently have to collaborate with on time-consuming processes such as testing.	Average team response ≤ 2.5
Q20–FC5	R16	When you spend money on the project, make sure you properly record everything related to the purchase.	Majority response = On budget
Q21–FC2/Q21–FC7/ Q21–FC8	R17	Make sure you are not rushing with your reasoning behind design decisions, safety decisions, or testing.	Majority response = Ahead of schedule
Q22–FC5	R18	If you think you are not satisfying technical requirements well, try to better record what you are doing and how it relates to the technical objectives.	Majority response = Less requirements satisfied
Q23		N/A	N/A
Q24		N/A	N/A
Q25		N/A	N/A

Table 29 continued

<i>Correlation Matrix Cell</i>		<i>Feedback Statement</i>	<i>Rule</i>
Q26– FC1/Q26–FC6	R19	When you solve a problem that came up, discuss as a team how your solution affects the rest of the design.	Majority response = Yes
Q27–FC8	R20	When you discuss how to prevent new risks in your project, do so in a structured and thorough way to properly account for all the things that could go wrong.	Majority response = Yes
Q28–FC1	R21	If a previous problem you had fixed is recurring, think about how other parts of the design have an impact on the problem and find a solution that addresses the root causes.	Majority response = Yes
Q29–FC1	R22	Come up with a thorough way for the team to report updates or changes for your designs, and do so every time, even if it seems inconvenient.	Majority response = No
Q30–FC7	R23	When there are a lot of outputs for the project in a week, put extra effort to thoroughly test all necessary components related to these new outputs.	Average team response ≥ 2.5
Q31	N/A		N/A
Q32	N/A		N/A
Q33	N/A		N/A
Q34–FC5	R24	Consider allocating more time each week to document your project work and to think about how you can manage risk (technical, budget, or schedule) for your project.	Average team response $\leq 50\%$
Q35– FC2/Q35– FC4/Q35– FC6//Q35– FC8	R25	During unscheduled team meetings retain your formal processes when it comes to communicating with everyone and justifying on your actions. Working outside class time is a good opportunity to get involved with tasks that you are less familiar with.	Average team response ≥ 2
Q36–FC1	R26	Think about whether you need to purchase any tools or equipment that could facilitate your design.	Average team response = 0
Q36–FC5	R27	When purchasing new tools or equipment, keep detailed records about where and how the purchased items are going to be used in the project.	Average team response ≥ 2
Q37	N/A		N/A
Q38	N/A		N/A

Table 29 continued

<i>Correlation Matrix Cell</i>		<i>Feedback Statement</i>	<i>Rule</i>
Q39– FC5/Q39-FC8	R28	When considering risks to your project, make sure that you are taking into account implications that are of technical nature (such as a component that does not follow your requirements and could disrupt the system).	Majority response = Budget
Q40		N/A	N/A
Q41– FC2/Q41– FC5/Q41-FC8	R29	Make sure you discuss as a team and in detail the implications your design decisions have, especially when it comes to safety features, and make sure everyone understands the justification behind them.	Majority response = Yes
Q42– FC1/Q42– FC6/Q42–FC8	R30	When having arguments about a topic related to the project with the team, focus on the more complex aspects that benefit from multiple perspectives such as identifying the project risks.	Majority response = Yes
Q43		N/A	N/A
Q44		N/A	N/A
Q45–FC2	R31	When thinking about the proper design for a component, talk to other teams that may have dealt with a similar problem to get ideas or learn the risks of using such a component.	Majority response = No
Q46–FC4	R32	Your involvement with this project is an opportunity to learn about many new topics, so make an effort to get involved with tasks you may not be familiar with.	Majority response = No
Q47– FC2/Q47–FC6	R33	Make a conscious effort to discuss with your team as a group why you are doing certain things, even if they appear trivial.	Majority response = No
Q48– FC5/Q48– FC10	R34	When making decisions on how to proceed with your project, do not reinvent the wheel if not necessary, but use readily accessible solutions. Pre-existing solutions can be easier for you to record, and have likely been evaluated previously as to how they impact your system.	Majority response = Think through
Q49– FC5/Q49–FC8	R35	Once you have some experience with proper reporting of your progress and thinking about risks for your projects, you may find yourself skipping steps during these processes due to overconfidence. Make a conscious effort to always have complete reports and risk analyses as they often can lead to problems later, if they are not done properly.	Average team response $\geq 50\%$

Table 30 shows a summary of the feedback statements provided to the project teams during experiment II.

Table 30: Summary of feedback statements provided to each of the project teams during experiment II, for all weeks. Each team received three feedback statements from a pool of recommendations dependent on the treatment group they were a part of. The statements were repeated if the team did not provide new answers for a given week.

	Project	Treatment Grp	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
Aerospace #1	Project 1	TF	R11/R22/R25	R1/R13/R29	R1/R13/R29	R19/R25/R33	R19/R25/R33	R19/R25/R33	R19/R25/R33	R19/R25/R33
	Project 2	GF	R14/R24/R27	R17/R24/R31	R16/R24/R31	R7/R12/R17	R15/R25/R27	R5/R17/R21	R8/R11/R17	R8/R11/R17
	Project 3	TF	R7/R9/R13	R9/R16/R25	R4/R14/R25	R15/R25/R35	R9/R14/R27	R2/R7/R14	R10/R13/R19	R10/R13/R19
	Project 4	GF	R6/R19/R27	R6/R19/R32	R6/R19/R32	R6/R13/R27	R6/R13/R27	R1/R5/R10	R3/R8/R12	R3/R8/R12
	Project 5	GF	R8/R18/R22	R8/R22/R29	R5/R11/R21	R7/R21/R27	R7/R21/R27	R13/R21/R25	R13/R21/R25	R13/R21/R25
	Project 6	TF	R1/R20/R28	R5/R14/R28	R5/R14/R28	R2/R13/R29	R11/R22/R25	R7/R16/R19	R7/R16/R19	R7/R16/R19
	Project 7	GF	R5/R19/R24	R5/R19/R24	R5/R19/R24	R21/R32/R33	R21/R32/R33	R10/R17/R20	R12/R24/R27	R12/R24/R27
	Project 8	TF	R9/R23/R35	R6/R14/R33	R6/R14/R33	R11/R16/R23	R11/R16/R23	R9/R23/R30	R13/R16/R26	R13/R16/R26
Aerospace #2	Project 9	GF	R13/R18/R28	R12/R18/R27	R12/R18/R27	R12/R18/R27	R3/R15/R28	R3/R15/R28	R5/R19/27	R10/R16/R25
	Project 10	TF	R11/R19/R24	R7/R19/R28	R15/R23/R32	R15/R23/R32	R22/R29/R30	R13/R19/R24	R13/R19/R24	R13/R19/R24
	Project 11	TF	R9/R16/R35	R16/R25/R30	R8/R9/R16	R14/R24/R32	R14/R24/R32	R14/R21/R30	R2/R9/R21	R7/R20/R35
	Project 12	GF	R3/R21/R32	R10/R17/R28	R14/R30/R34	R6/R15/R17	R15/R17/R28	R3/R18/R28	R12/R19/R21	R5/R10/R22
Aerospace #3	Project 13	GF	R20/R28/R34	R20/R28/R34	R13/R18/R25	R13/R18/R25	R10/R25/R34	R14/R15/R24	R12/R24/R26	R12/R24/R26
	Project 14	TF	R13/R23/R31	R13/R23/R31	R13/R23/R31	R4/R16/R31	R4/R16/R31	R14/R23/R27	R14/R23/R27	R14/R23/R27

5.5 Feedback Effectiveness Evaluation

As described earlier, I assigned the student teams to two treatment groups based on the feedback they received: targeted feedback (“TF”) or non-targeted feedback (“NTF”). I conducted two types of analyses to test whether the targeted feedback was helpful at reducing failures in student projects: two quantitative proportions statistical tests and a qualitative evaluation from student responses. I conducted the statistical tests using the project failure rates as reported by the instructors at the end of the semester. I included the qualitative measures as three separate questions in the student survey during Experiment II.

For the quantitative tests, I used two Barnard’s exact tests, similarly to Section 3.1. For the first test, I considered the three separate tests for each of the failure metrics (budget, schedule, requirements) separately. In this first case, each statistical test answered whether the feedback improved the student projects in terms of a specific failure metric. For the second test, I assumed that the failure metrics are independent and can be considered together as “failure metrics”. The assumption is common in literature (e.g., see Nan and Harter, 2009) as those metrics are considered independent measures of project success. In this second case, the statistical test answered whether the targeted feedback improved the student projects in terms of any failure metric. The motivation behind running both tests was because of the small sample of projects potentially limiting the power of the separate tests. For the budget test only, the instructor of the “Aerospace #1” course was not able not provide any budget data due to the nature of the projects. Table 31 shows the instructor project evaluations at the end of the semester for each project.

Table 31: Instructor evaluations at the end of semester during Experiment II. Instructors provided failure metrics for the 14 student projects. The project numbers do not correspond to the actual project team names for confidentiality purposes. One course was not viable for budget evaluation. “1” corresponds to failure and “0” to success.

Course	Project #	Treatment Group	Budget Failure	Schedule Failure	Requirements Failure
Aerospace #1	1	TF	—	1	0
	2	NTF	—	1	0
	3	TF	—	0	0
	4	NTF	—	1	0
	5	NTF	—	0	1
	6	TF	—	1	1
	7	NTF	—	1	0
	8	TF	—	1	0
Aerospace #2	9	NTF	0	1	0
	10	TF	0	0	0
	11	TF	1	1	1
	12	NTF	0	1	0
Aerospace #3	13	NTF	0	0	0
	14	TF	1	0	0

Using the instructors' assessment of each project at the end of the semester (Table 31), I computed the sample estimate of failure proportion for each metric and group as follows:

$$\hat{F}_{(TF),j} = \frac{\sum_{k=1}^{n_1} failure_{i,k}}{n_1} \quad (Equation 19)$$

Where j is one of the three failure metrics, n_1 is the number of student projects in the targeted feedback group, and $failure_{i,k}$ is a binary variable that is equal to 1 if project team k failed in terms of metric j at the end of the semester or 0 if not.

For the non-targeted feedback group, the sample estimate of failure proportion is defined similarly:

$$\hat{F}_{(NTF),j} = \frac{\sum_{k=1}^{n_2} failure_{i,k}}{n_2} \quad (Equation 20)$$

Where j is one of the three failure metrics, n_2 is the number of student projects in the non-targeted feedback group, and $failure_{i,k}$ is a binary variable that is equal to 1 if project team k failed in terms of metric j at the end of the semester or 0 if not.

Based on the previous definitions and data, the estimated failure proportions are:

Table 32: Estimated sample failure proportions for budget, schedule, and requirements based on the instructor evaluation at the end of Experiment II.

<i>Treatment Group</i>	<i>Budget</i>	<i>Schedule</i>	<i>Requirements</i>	<i>Combined failure metrics</i>
TF	$\hat{F}_{(TF),1} = 2/3$	$\hat{F}_{(TF),2} = 4/7$	$\hat{F}_{(TF),3} = 2/7$	$\hat{F}_{(TF),comb} = 8/17$
NTF	$\hat{F}_{(NTF),1} = 0/3$	$\hat{F}_{(NTF),2} = 5/7$	$\hat{F}_{(NTF),3} = 1/7$	$\hat{F}_{(NTF),comb} = 6/17$

Based on these results, the targeted feedback teams appear to have performed worse in terms of budget and requirements compared to the non-targeted feedback teams. For the schedule metric, targeted feedback teams did slightly better (1 more successful project compared to non-targeted feedback teams). There are at least two possible reasons for these observations:

1. The students of targeted feedback teams may not have known how to turn the feedback into an action that would positively impact their project (e.g., for R1, they may not have known *how* to communicate better).
2. Although the non-targeted feedback statements did not necessarily address a team's weak areas, they may nevertheless have been more effective because they addressed a wider range of potential failure causes, resulting in better team overall performance.

To statistically quantify the results, I used Barnard's exact test with the null hypothesis that the failure rate of groups that received targeted feedback in terms of metric j is equal to or more than the failure rate of groups that received non-targeted feedback in terms of metric j . I selected a significance level α equal to 0.05 and I repeated the test for all three independent metrics.

$$\begin{aligned}
H_0: F_{TF,j} &\geq F_{NTF,j} \\
H_a: F_{TF,j} &< F_{NTF,j}
\end{aligned}
\tag{Equation 21}$$

Overall, the failure rates show that the targeted feedback does not reduce the failure occurrences in terms of any metrics, compared to the non-targeted feedback.

Table 33: Barnard's statistical test results for the targeted feedback. The statistical test suggests that the targeted feedback statements do not reduce the occurrence of failures in student projects, compared to the non-targeted feedback statements.

<i>Failure metric</i>	$\hat{F}_{(TF),j}$	$\hat{F}_{(NTF),j}$	<i>Barnard's test one-tailed p-value</i>	<i>H0 rejected?</i>
Budget	2/3	0/3	0.958367742	No
Schedule	4/7	5/7	0.288499581	No
Requirements	2/7	1/7	0.742586144	No
Combined	8/17	6/17	0.757077341	No

Apart from the statistical test, I also added three questions to the student survey, to gauge: whether the students actually change their behavior due to the feedback, how they receive it, and whether they think it actually helps them. The added questions are shown below (Table 34).

Table 34: The three additional questions that were part of the student survey during Experiment II to gauge how they received the feedback.

Feedback Evaluation	
F1	Did you do anything differently during the past week because of the feedback we gave you for your project? (<i>Yes/No</i>)
F2	How helpful do you think the feedback was? (<i>Likert scale answer: Not helpful at all (1) to Very helpful (5)</i>)
F3	How likely do you think it is for the feedback to improve your project's performance in any way? (<i>0–25%, 25–50%, 50–75%, 75–100%</i>)

Students in the teams that received targeted feedback appear to have responded better to the feedback by learning something and changing how they go about their project. 68% of the responses from the targeted feedback teams show that students did something differently, compared to 34.8% from the teams that received non-targeted feedback (Figure 16). Regarding the question on how helpful the students thought the statements were to them, teams that received non-targeted feedback had normally distributed responses to the question, with the majority (39.1%) rating the feedback as a 3 on a scale of 1–5. On the contrary, teams that received targeted feedback have skewed responses towards higher ratings, with “very helpful” being the most frequent answer (30%) (Figure 17). In the last question, both treatment groups show similar number of responses in the categories covering 25-75% of positive impact. Most of the students saying the feedback is “>75%” likely to positively impact their project performance come from targeted feedback group responses.

Overall, the evaluations from the students indicate that the targeted feedback is better received by the students compared to non-targeted feedback: they are more likely to change their behavior, they find it more helpful, and they believe it can help them with their project performance.

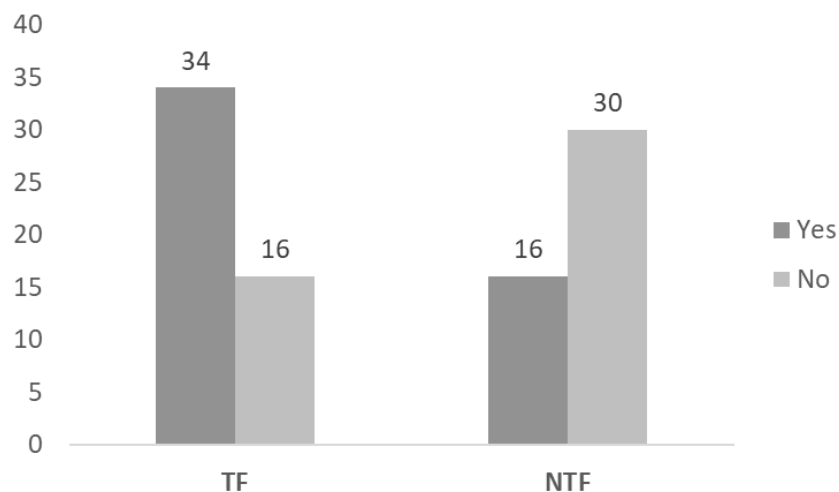


Figure 16: 68% of the responses (34 out of 50) from teams that received targeted feedback show that the students changed their behavior, compared to 34.8% of the responses (16 out of 46) from teams that received the non-targeted feedback.

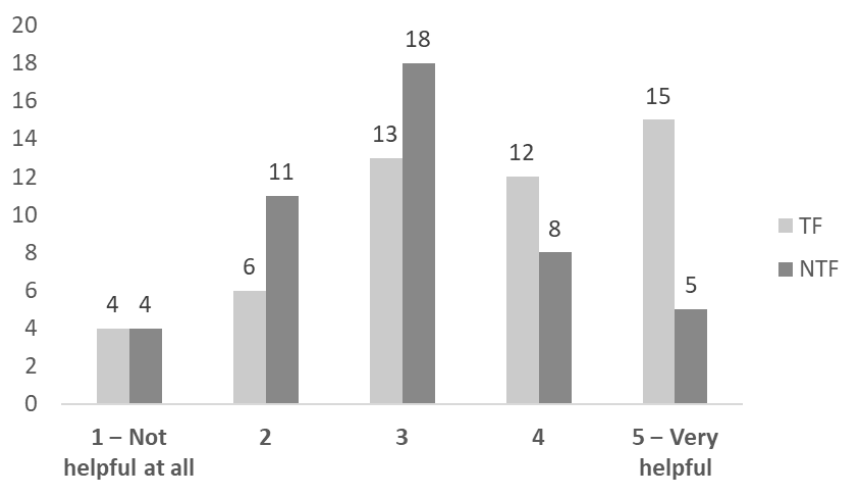


Figure 17: Teams that received non-targeted feedback rated, most of the time, the feedback statements as moderately helpful, with more responses (15) scoring it as 1 or 2 than 4 or 5 (13). Responses from teams that received targeted feedback are towards higher ratings, with most responses (15 out of 50) rating the statements as very helpful.

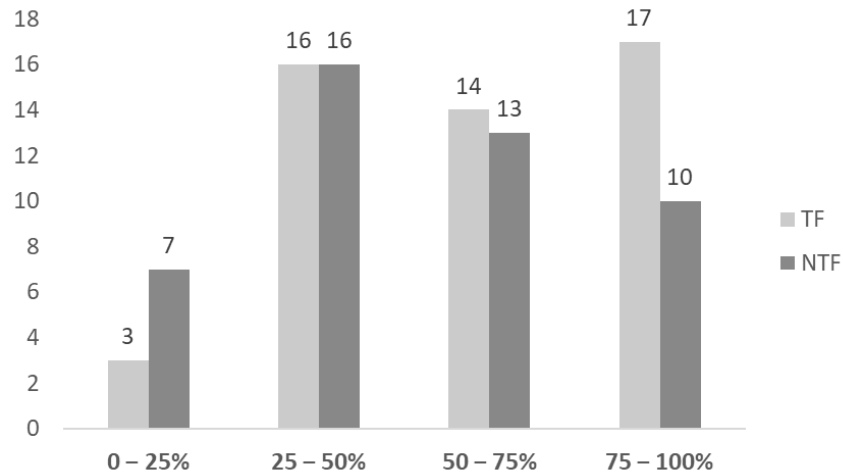


Figure 18: Both treatment groups thought the feedback would have some positive impact on their project, possibly due to the positive nature of the statements. The responses from the targeted feedback teams had a larger representation in the >75% chance of positive impact answer and smaller representation in the <25% positive impact answer, compared to the non-targeted treatment group. The “TF” group had 4 more answers than the “NTF” group, which contributes to the observation.

Synthesizing the results between the quantitative and qualitative comparison of the feedback, I found that:

1. Students who received the targeted feedback statements said they were more likely to change their behavior. However, the overall project success rates did not improve, suggesting that the students either did not make any changes or whatever changes they made were ineffective, perhaps because they did not know *how* to change their behavior appropriately.
2. Students said the targeted feedback was more helpful than the non-targeted feedback, but the project success rates indicate that they were only able to improve in terms of the schedule metric.
3. More students in the targeted feedback group said that the feedback statements would have a positive impact on their projects, but end-of-semester success rates do not agree with these responses. It is possible that the feedback contributes positively to the student projects, but is not enough to have an impact on project success.

6. COMPARISON OF CROWD SIGNALS BETWEEN EXPERIMENTS I AND II

This chapter shows some descriptive statistics for the student responses to the questions that collect the crowd signals, and compares the student answers between Experiment I and II. The main difference between the two experiments was the presence of feedback (Experiment II) or not (Experiment I). Therefore, I wanted to further investigate any differences in the student responses due to the presence of feedback. The second experiment also included fewer questions because some of the original crowd signals were removed during the model reduction process (see Section 4.3).

Figures 19 to 29 show descriptive statistics for the responses to the *Performance* questions (Q1 to Q8).

The majority of the 74 students who participated in the first experiment had previous experience ranging from 1–4 engineering projects before joining the courses from which I collected data (Figure 19). Experiences include projects from other coursework, internships, and extracurricular activities. The projects I monitored come from senior-level courses, and so it is expected that some students will have more experience than others. Q1 was not included in Experiment II.

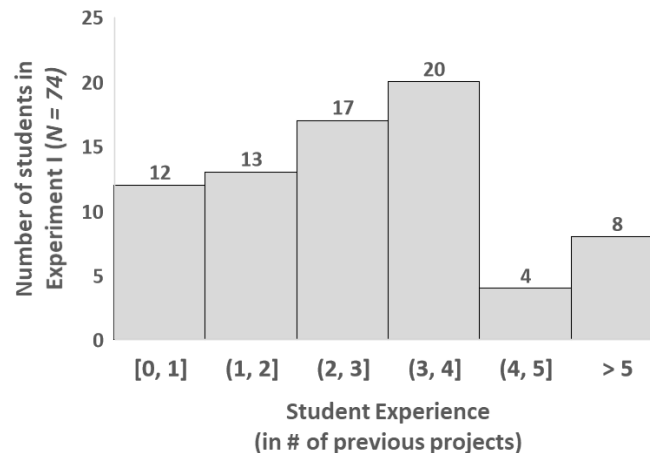


Figure 19: Experience level of the $N = 74$ students that have participated in experiment I. I measured experience based on the number of engineering projects the students have been part of in the past. Most students had previous experience from 1–4 projects.

I measured *proactivity* by asking how many times the students got involved with other tasks outside their immediate responsibility (Q2) and present here the averaged the responses per individual student. For Experiment I, 34 students attempted to get involved with other tasks a maximum of once in a given week, and 27 of them attempted to get involved twice (Figure 20). Very few made more than three attempts. There is a similar result for the *proactivity* measure for Experiment II, however a larger percentage of students attempted to get involved between 1 to 2 times per week (Figure 21). Getting involved 1 time per week may appear quite low, but considering that students have their own project tasks to complete and they only meet for a total of about 2.5 hours of laboratory work every week, there may not be too many opportunities to get involved with tasks outside their immediate responsibilities.

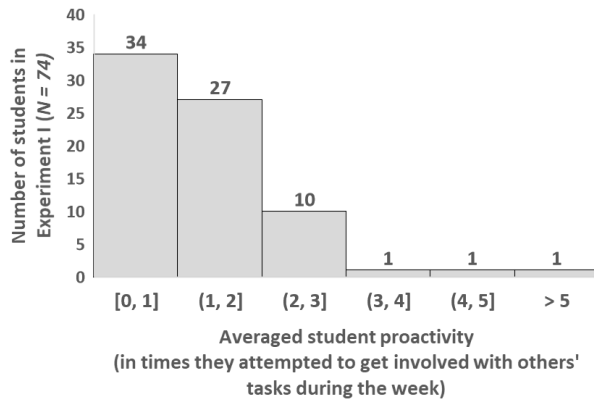


Figure 20: Most of the $N = 74$ students that participated in Experiment I, got involved, on average, at most one time with tasks outside their immediate responsibilities.

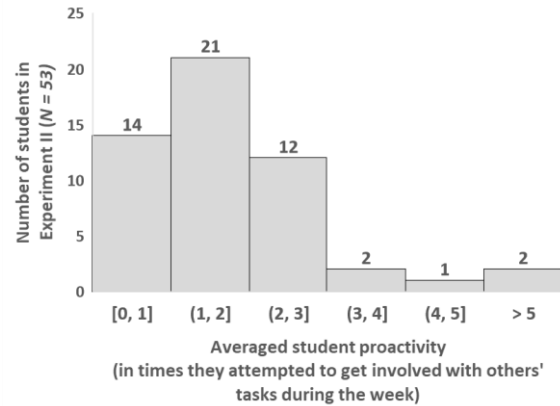


Figure 21: Most of the $N = 53$ students that participated in Experiment II, got involved, on average, between 1 to 2 times with tasks outside their immediate responsibilities.

Figures 22 to 24 compare the responses for question Q3 (*stress level*), Q4 (*coordination*), and Q5 (*team impact*) between the two experiments. For these statistics, because the answers can change weekly based on the students' and project activities, I did not average by individual student. Instead, I collectively averaged all the responses by the total number of observations ($N = 304$ for Experiment I and $N = 132$ for Experiment II) to find the answers that are most frequent.

For Q3, the majority of responses showed that students were sometimes unable to focus on their projects for both experiments. 40% of students in Experiment II said they “rarely” had problems focusing, compared to 34% of students in Experiment I.

For Q4, most responses showed that students interacted at least sometimes with their team when working on separate tasks, which suggests that there was some level of coordination between them. Only 1 response in Experiment II indicated no interacted with the team. A larger percentage of responses in Experiment II showed that students “very often” interacted with their team members when working separately (51%, compared to 28% Experiment I), possibly due to the COVID-19 restrictions, forcing students to meet virtually a lot more frequent to make progress even when working in separate physical locations.

Most responses to Q5 showed that students made meaningful progress towards their project goals during a typical week. The teams met two or three times during the week and had to make progress every time to meet their milestones. The majority of the students said they very often made a meaningful progress (44% for Experiment II compared to 28% for Experiment I). The responses to Q4 and Q5 indicate that feedback helps students in interacting with each other and making meaningful progress more often.

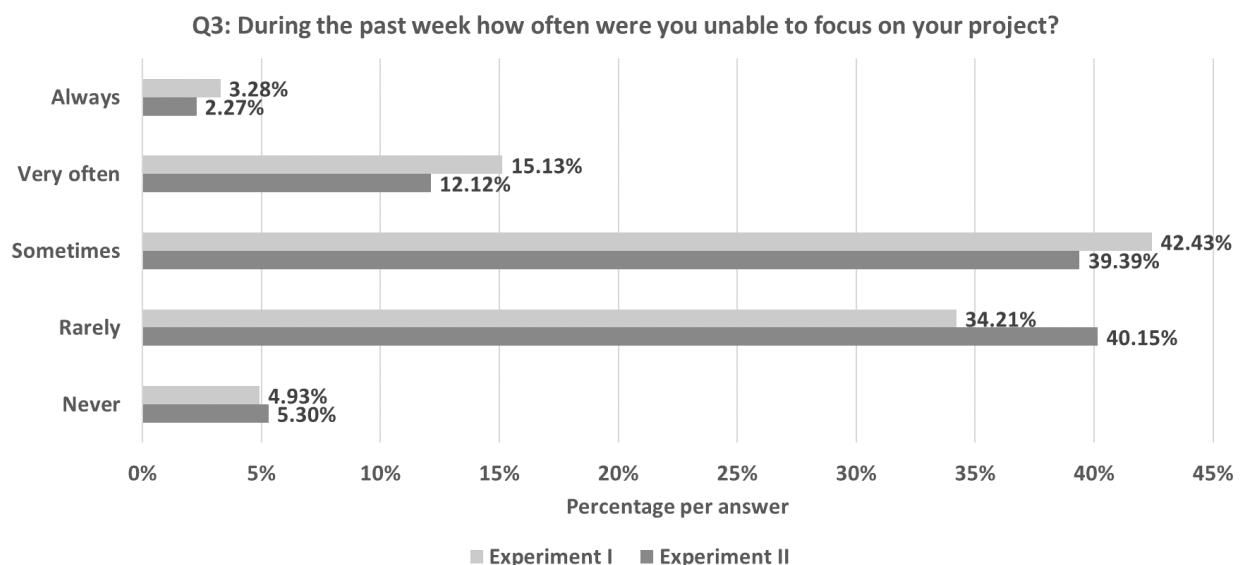


Figure 22: Statistics of the student responses to Q3. The majority of responses showed that students were sometimes unable to focus on their projects.

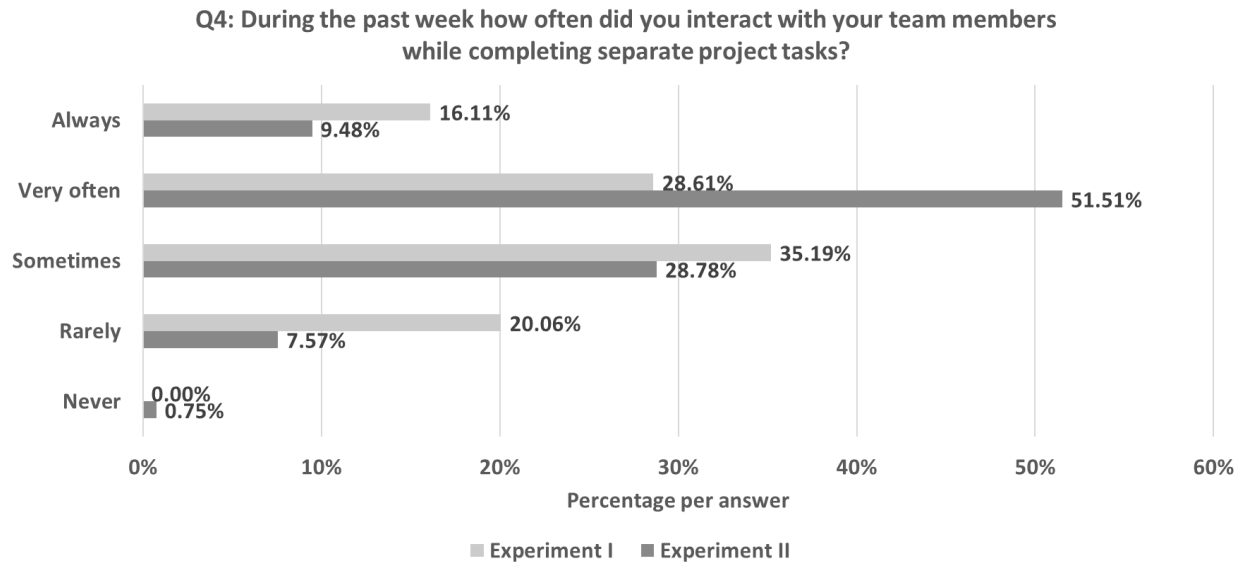


Figure 23: Statistics of the student responses to Q4. For both experiments most showed that there was some level of coordination even when working on separate tasks, but students in Experiment II interacted more often with each other.

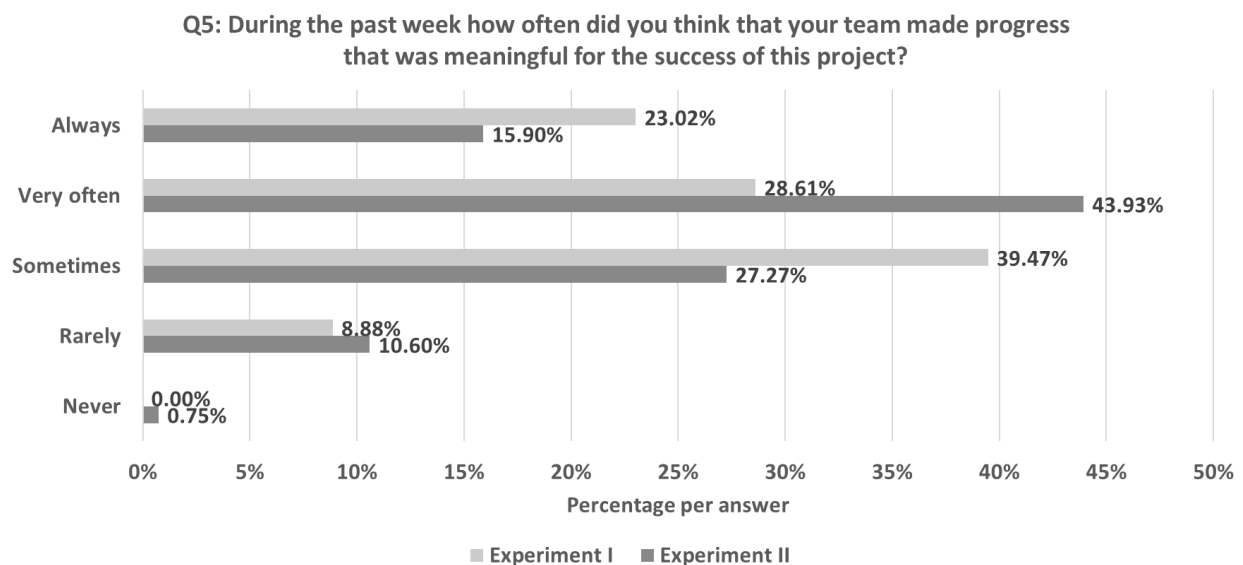


Figure 24: Statistics of the student responses to Q5. For both experiments most responses showed there was some level of meaningful progress during a typical week, but students in Experiment II made such progress more frequently.

Figure 25 and Figure 26 show what percentage of their team's activities the students knew. During Experiment I, most students knew what at least half of their team members were working on. Considering that most teams include 3–5 people with overlapping tasks, this percentage indicates that the students were mostly familiar with their team members' work, excluding one or two people, who may be the ones with separate and independent responsibilities (e.g., manufacturing a component). Experiment II responses indicated that students knew less about their team members' tasks, compared to Experiment I. Safety measures imposed due to the COVID-19 pandemic may have been responsible for that, given students had to follow capacity limits in their workspaces so they could not work in the same room as much.

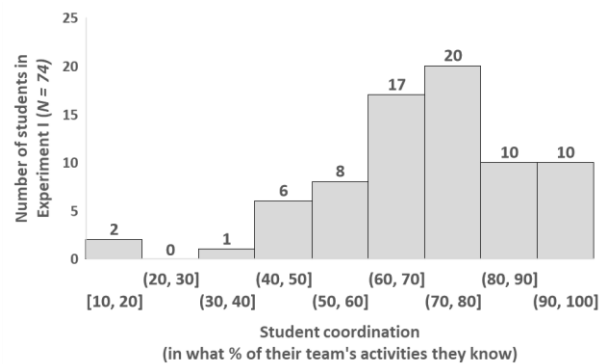


Figure 25: Of the $N = 74$ students who participated in Experiment I, most knew at least what half of their team members were working on.

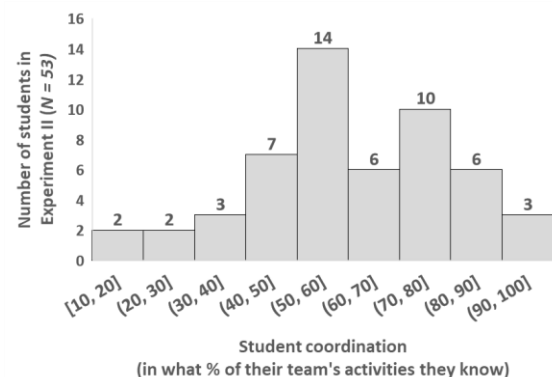


Figure 26: During experiment II, students did not know as much about their team members' activities as during Experiment I. Hybrid learning and safety measures due to the COVID-19 pandemic may be responsible for that, given students have to follow capacity limits in their workspaces so they could not work in the same room as much.

Q7 asked the students to rate how much freedom the instructors gave them when completing project tasks. The most frequent response during both experiments was “much freedom” (35.85% during experiment I and 43.93% for experiment II (Figure 27)).

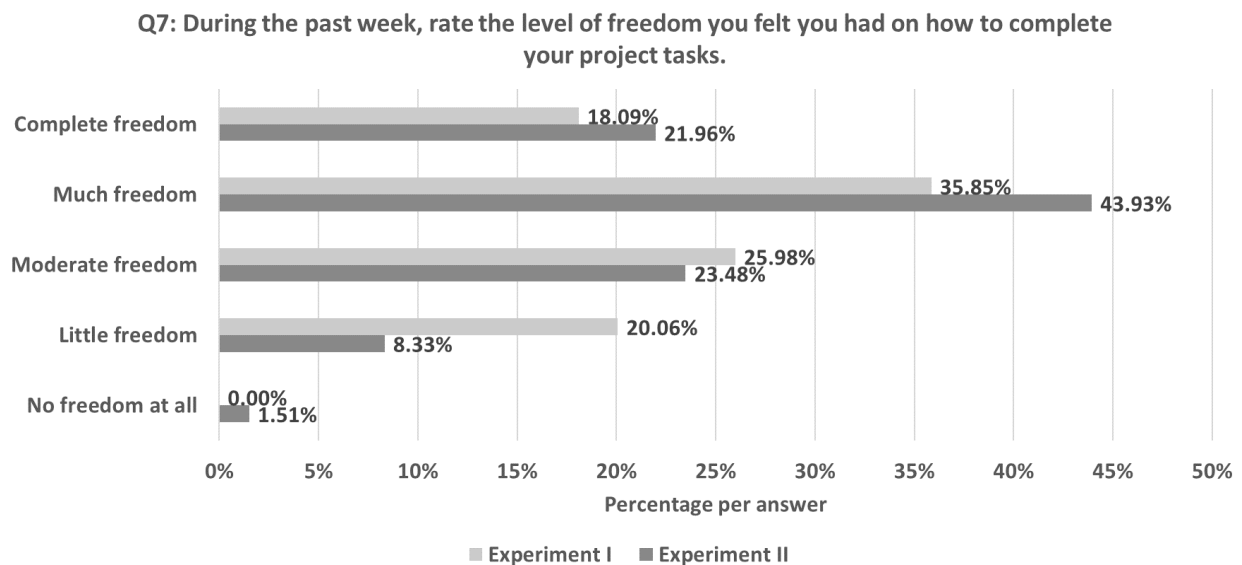


Figure 27: Statistics of the responses to Q7. Only two responses during Experiment II indicated students had no freedom at all when completing a project task. More than 75% of responses showed that the instructors for the courses in both experiments gave students at least moderate freedom on how to complete their project objectives.

Figure 28 and Figure 29 show for each student, the average chance of success they thought they had if there were no instructor oversight for the remainder of the semester. If a student gave a high chance to their team, that would be one indication of a team member who had confidence in the team’s capabilities to work autonomously. For Experiment I, students gave varying chances of success if their teams would have to complete the projects without supervision for the remaining of the semester. For experiment II, responses appear to collect around confident students (>70% chance of success without instructor) and less confident students (30–60% chance of success without instructor).

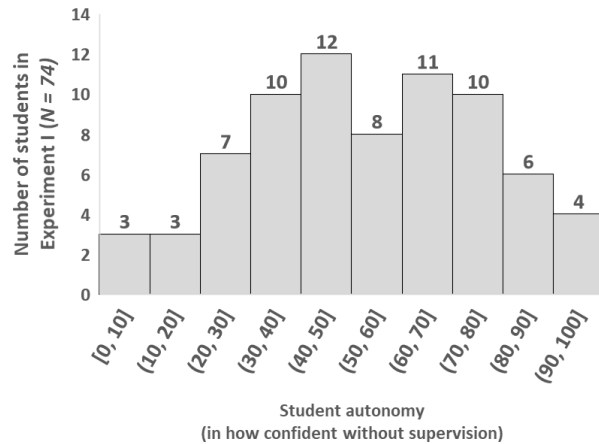


Figure 28: During Experiment I, students gave varying chances of success if their teams would have to complete the projects without supervision for the remaining of the semester. The responses to this question likely depend on the project phase, the student's confidence in their own and their team's capabilities, and overall knowledge about the project's next steps.

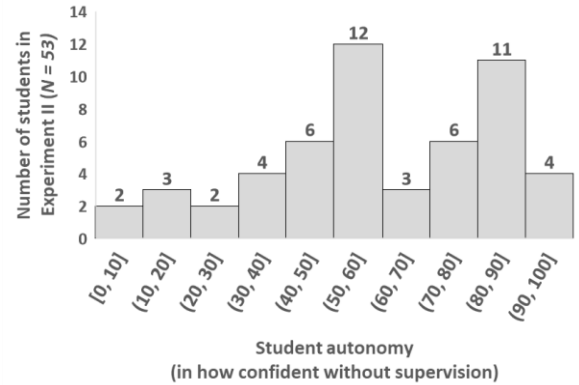


Figure 29: During Experiment II, the majority of the responses from students appear to get grouped around two options: 1) students that think they need the instructor supervision and would only have 30–60% of success without it, and 2) students that think they could be successful (>70%) without the instructor.

For Q9 (“During the past week, which of the following attributes/adjectives relating to creativity do you feel apply to your team’s project work?”), I do not show any statistics as the responses include a number of adjectives that are associated with creative or uncreative designs from literature. I used the information when building the predictive models to classify *creativity*, which is one of the predictors, between three levels (low, medium, or high). I describe how I came up with the coding scheme in Table 9 (Chapter 3.2).

Figures 30 to 35 show descriptive statistics for the responses to the *CSF* questions (Q10 to Q14). For Experiment I and II, the percentage of tasks that can be performed independently of the rest of the project, indicating the level of modularity in the design of the project, shows a uniform profile (Figure 30 and Figure 31). The number of these modular tasks depends on the project phase and the current technical requirements that change week to week.

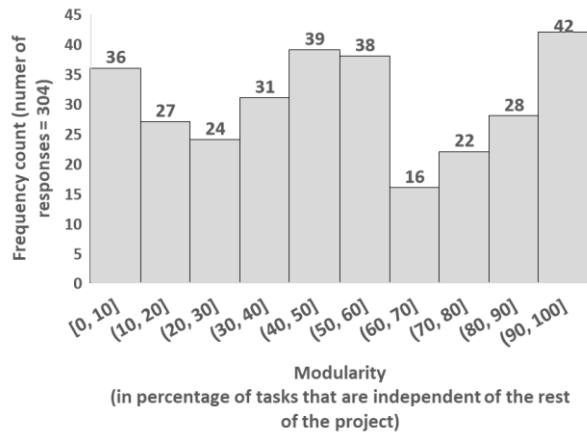


Figure 30: The percentage of tasks that can be performed independently of the rest of the project resembles a uniform profile for Experiment I. The degree of modularity in a project is likely dependent on project phase and current requirements that continuously change and get updated throughout the semester.

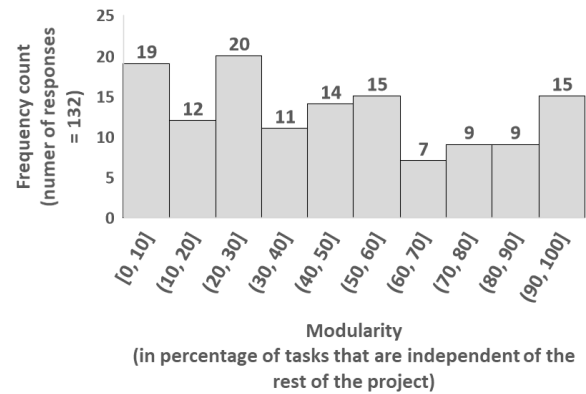


Figure 31: Similar to Experiment I, the percentage of tasks that can be performed independently of the rest of the project resembles a uniform profile for Experiment II.

Most of the responses (more than 70% for Experiment I, more than 87% for Experiment II) show that for the majority of projects, the objectives were at least moderately clear, which suggests that most of the time the students knew what the next steps were for the project (Figure 32). The instructors who managed the projects likely told the students exactly what the needs were for a specific task, which made the objectives quite clear.

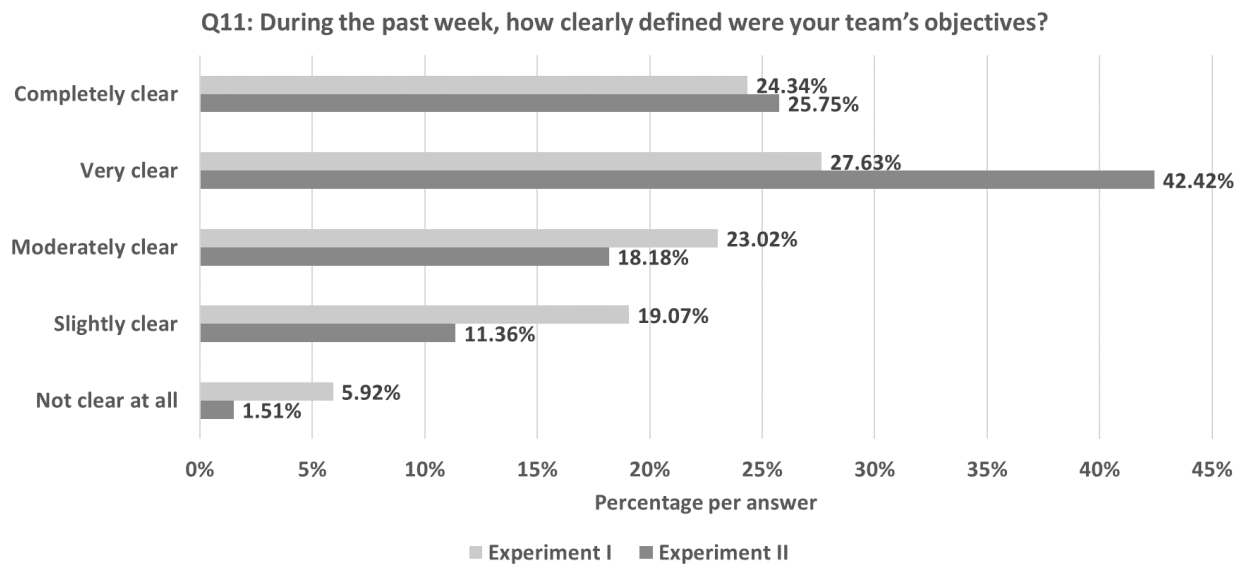


Figure 32: Statistics of the responses to Q11. More than 75% responses indicate that the project objectives were at least moderately clear for both experiments. The objectives were likely set by the instructor, who directly asked for a task to be completed and gave specific milestones for the students. Also, some teams may have set their own objectives, since they may have some freedom on how to complete the project tasks.

Responses to Q12 showed that the majority of students would definitely continue to work on the project with a completely new team (41% for Experiment I and 36% for Experiment II) if the rest of their team quit (Figure 33). Commitment to project success may be related to the quality of the courses and the learning opportunity during these projects.

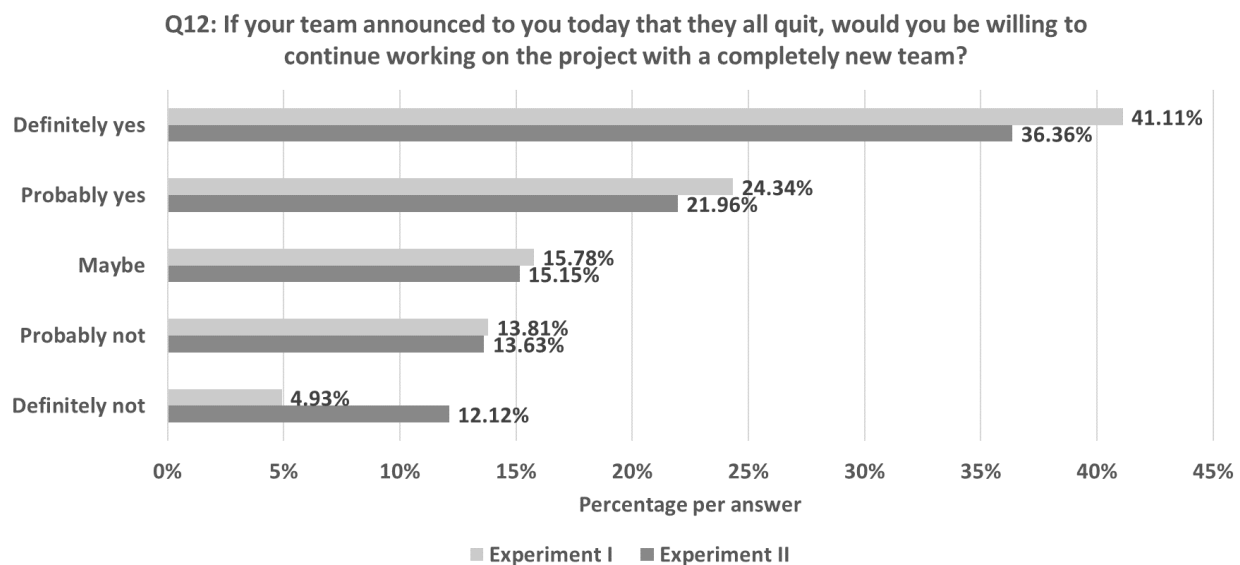


Figure 33: Statistics of the responses to Q12. More than 80% of responses indicate that students would at least consider working on the project with an entirely new team for both experiments. 12% of responses for Experiment II show that students would definitely abandon the project (compared to only 5% in Experiment I). The results are likely related to the quality of the courses, the value the students saw in participating, and how long they were part of the project.

For both experiments, more than 85% of responses indicated that students had at least moderate availability of resources to use during a given week (Figure 34). Such resources are often software for designing parts before manufacturing or running simulations, lab equipment and tools, and funds. Very few responses (only 1%) point to very low availability of resources, suggesting that students generally had access to the resources they needed to succeed.

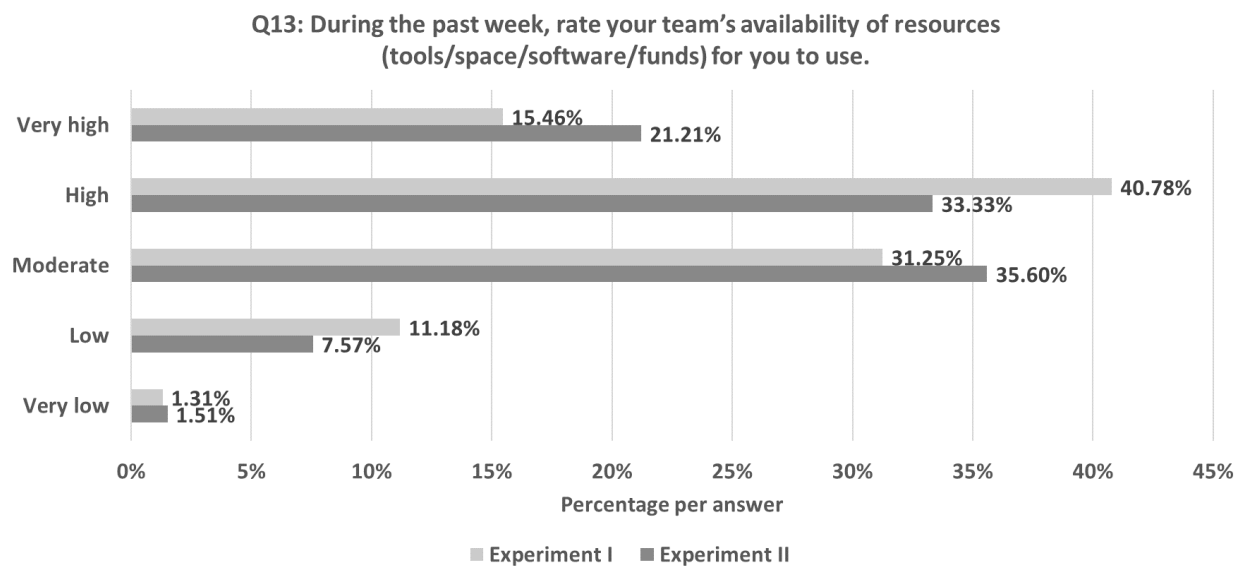


Figure 34: Statistics of the responses to Q13. More than 85% of responses indicate at least moderate availability of resources to use in a given week. The quality of the laboratories and campus resources are likely related to the response profile for this question. Very few responses communicated very low availability, perhaps due to a specific need or request for the project that was not met.

Lastly, 41.1% said that they noticed lack of communication with their team while working on their project during Experiment I (Figure 35). 6.9% of responses show that students always noticed a silent room, which can be concerning as it may be a sign of poor team cohesion and correlate with failure, but may also happen if students were working on their own. Q14 was not included in Experiment II.

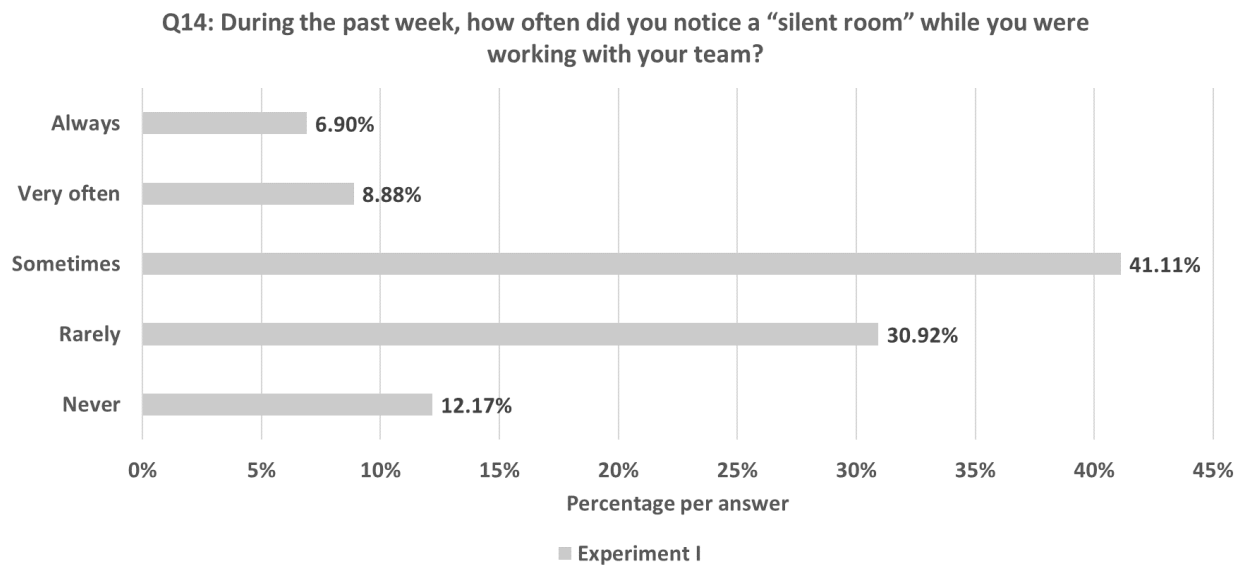


Figure 35: Statistics of the responses to Q14 (averaged by all $N = 304$ responses) during Experiment I. 41.1% of responses indicate that sometimes there is lack of communication while students work together. 6.9% said that they always noticed a silent room, which may be a sign of poor team cohesion. Poor communication is a frequent problem amongst teams in engineering projects, and student projects also confirm that.

Figures 36 to 40 summarize the responses to the *Individual Personality* questions (Q15 to Q19) for the two experiments.

Responses show that the majority of students rarely get frustrated with each other and their team (35.5% for Experiment I, 40.7% for Experiment II), closely followed by getting frustrated sometimes (30.2% for Experiment I, 25.7% for Experiment). 2–3% of responses come from students who said they were always frustrated during a particular week, which perhaps happens during the duration of a project because of disagreements, failure, or poor teamwork.

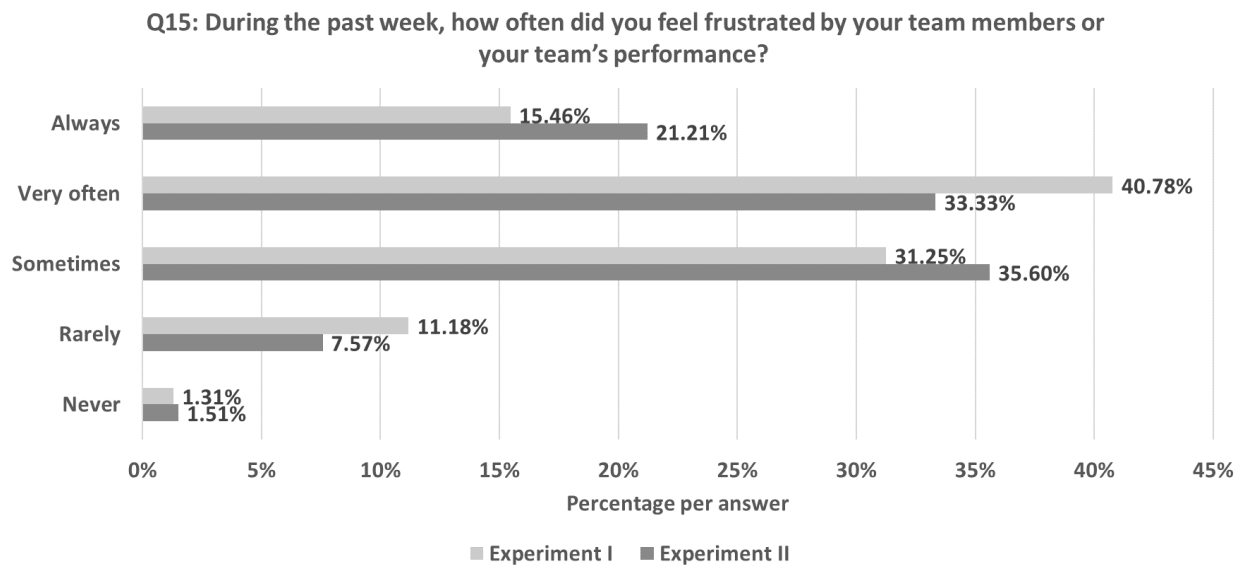


Figure 36: Statistics of the responses to Q15. The majority of students rarely get frustrated with each other and their team.

For Experiment I, 44.7% of responses show occasions where students were open to new ideas, either by suggesting them or by agreeing to someone else's. The response to the question (Q16) may change depending on the project activities for a given week. For example, if there were mundane tasks to complete then perhaps students did not discuss new ideas, but rather completed the work. For Experiment II, Q16 showed a more balanced profile of responses compared to Experiment I (each of the “rarely”, “sometimes”, “very often” options received close to 25% of responses).

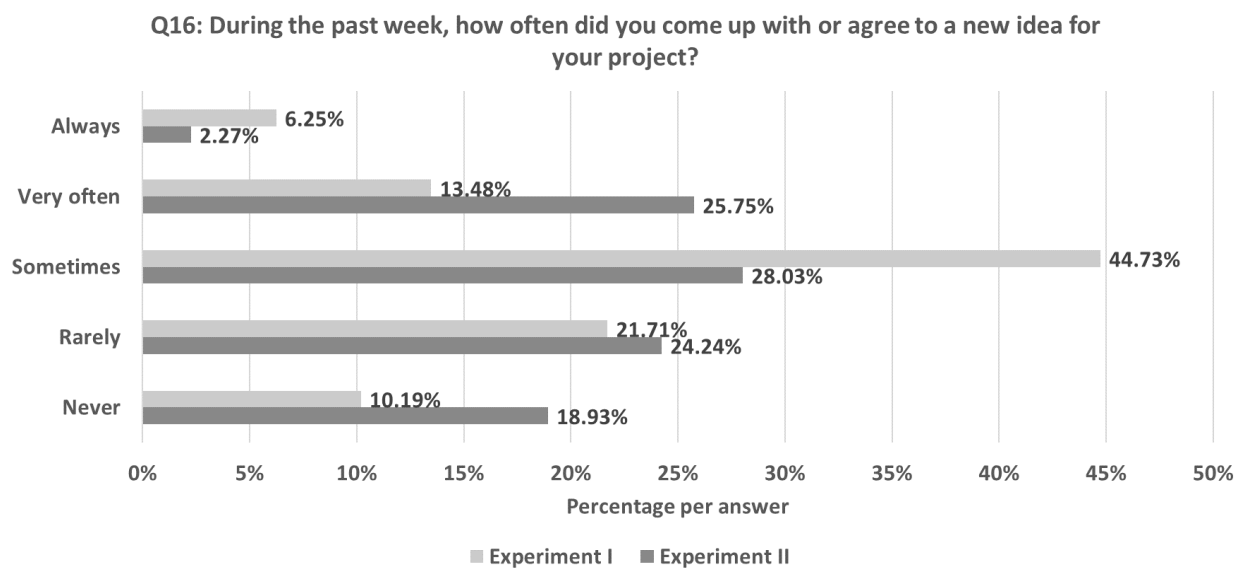


Figure 37: Statistics of the responses to Q16. For Experiment I, the majority of responses reflected that students “sometimes” come up with or agree to new ideas for the projects. For Experiment II, the responses showed a more balanced profile compared to Experiment I.

The majority of responses for Q17 (62% for Experiment I, 42% for Experiment II) showed that students would rarely or never skip or cancel an obligation or task in a given week, while some indicated they always did (9.8% for Experiment I, 1.5% for Experiment II). There are some circumstances that would force students to skip on all required activities, perhaps due to personal matters, and the frequency of these occurrences may be correlated with how well the team performs without the individual. During Experiment II, it is likely students would overall need to skip more activities due to reasons external to the team.

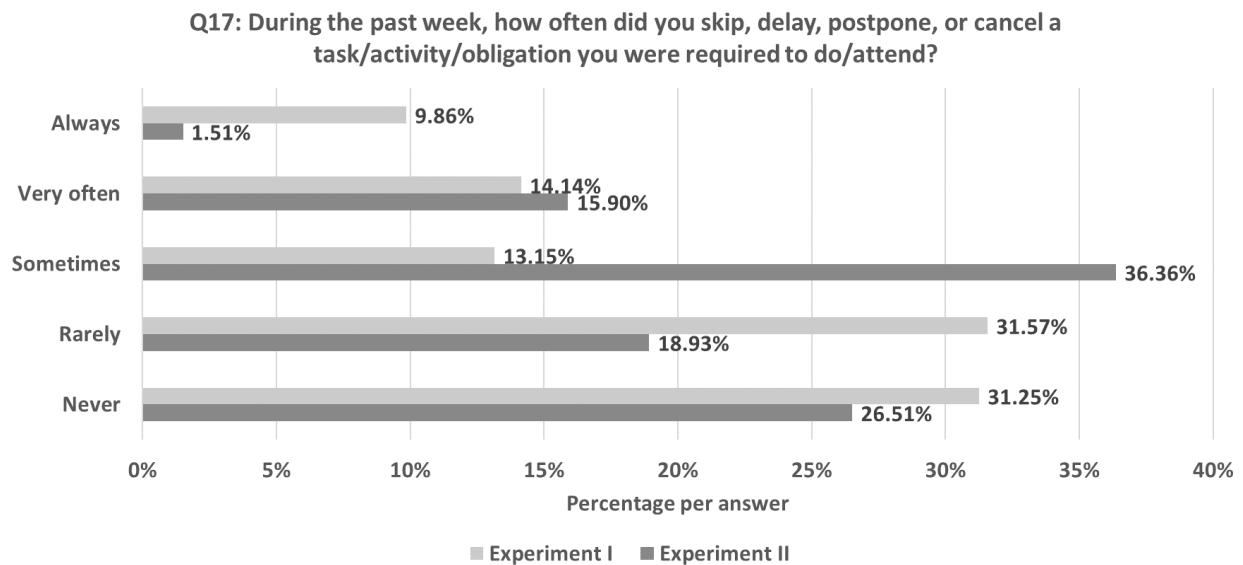


Figure 38: Statistics of the responses to Q17. The majority of students would rarely or never skip or cancel an obligation or task in a given week.

For Q18, 14.8% of responses for Experiment I and 30% for Experiment II show that students never felt that they were the center of attention, while 11.1% for Experiment I and 2% for Experiment II said they always did. The responses to this question likely have to do with the personality of the student, and whether their task or responsibility was the focus in a given week. Also, the COVID-19 restrictions during Experiment II perhaps have an impact on how the teams worked together, not allowing perhaps much room for one member to become the center of attention.

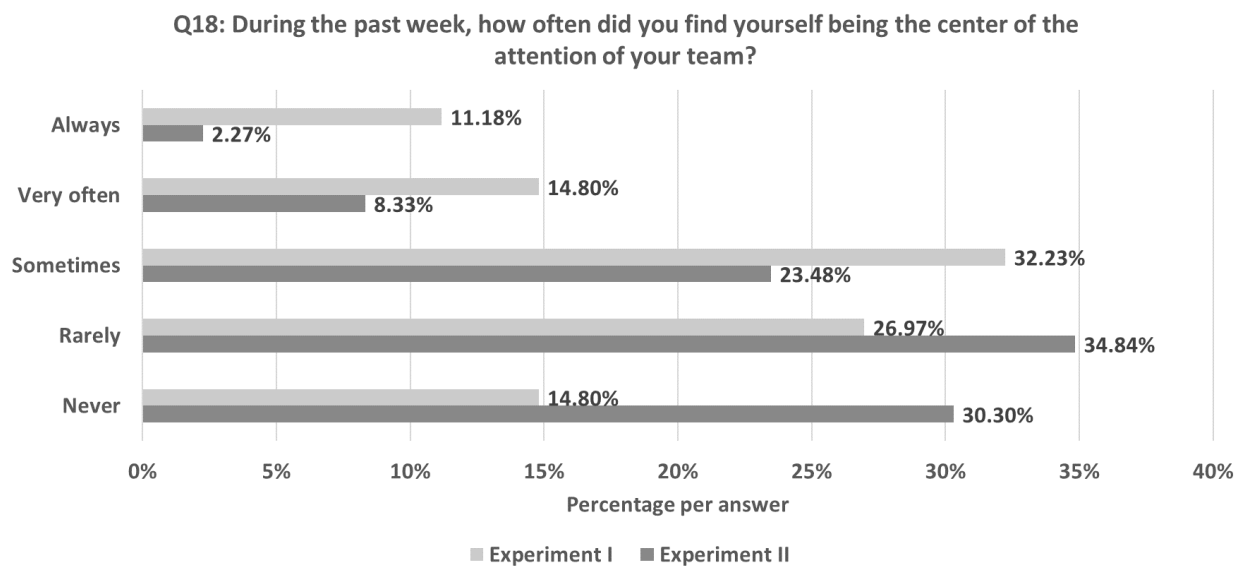


Figure 39: Statistics of the responses to Q18. A relatively small number of responses showed a student to always be the center of attention.

For Q19, 19.1% of responses in Experiment I and 13% in Experiment II that students never had one of their team members share important things about their life with them, while only 6.2% for Experiment I and 2% for Experiment II said they always did. Students sharing details about their lives may be an indication of a friendly working environment.

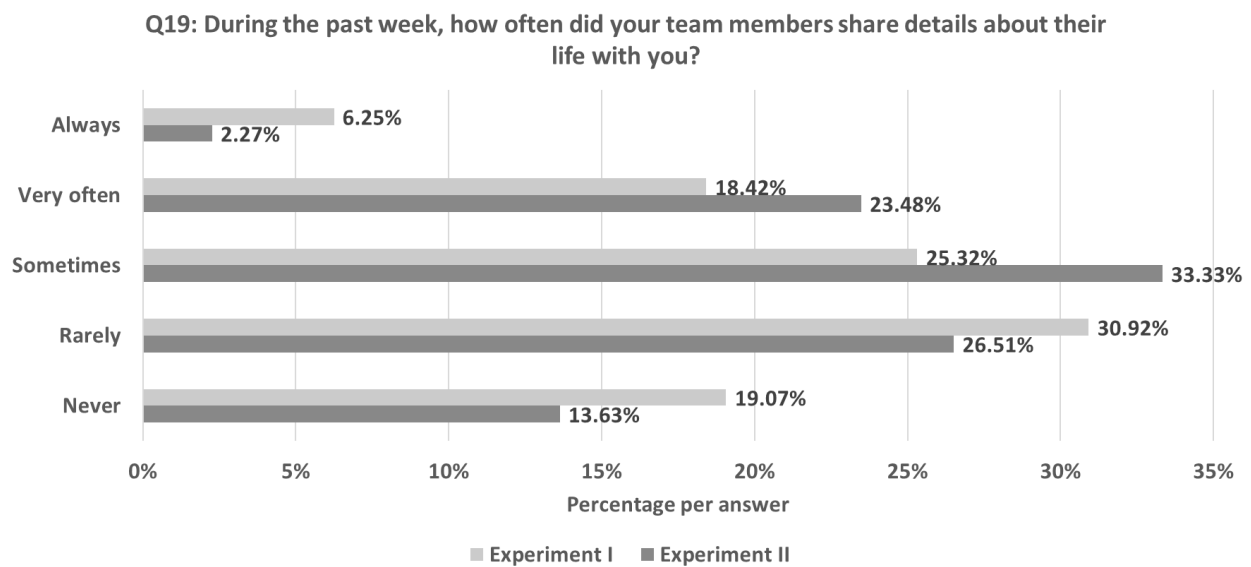


Figure 40: Statistics of the responses to Q19. 19.1% of responses in Experiment I and 13% in Experiment II that students never had one of their team members share important things about their life with them, while only 6.2% for Experiment I and 2% for Experiment II said they always did.

Figures 41 and 42 show histograms of student confidence in their spending estimates during Experiments I and II, respectively. Very few students admitted low confidence in their spending estimate. For Experiment I, most responses concentrated around 40–50% and 70–100%, perhaps representing students who were moderately confident and very confident respectively. For Experiment II, the confidence profile appears to be more uniform in the 40–90% range with a large increase in the 90–100% range indicating absolute confidence.

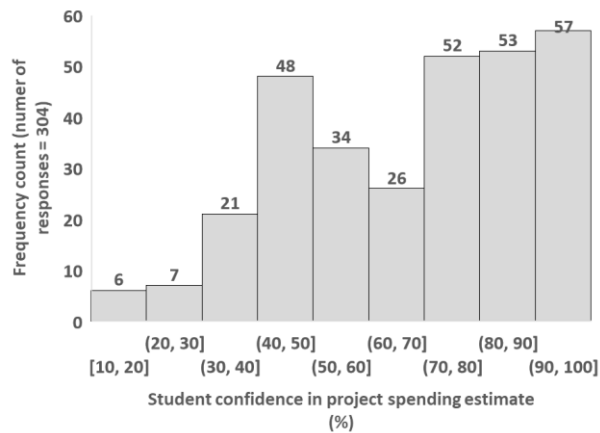


Figure 41: Student confidence in their spending estimate for Experiment I. Most responses concentrated around 40–50% and 70–100%.

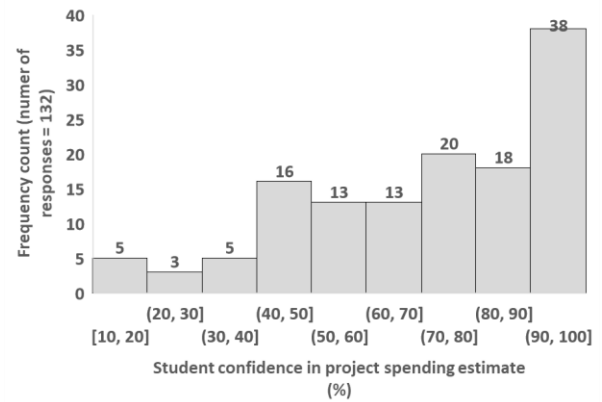


Figure 42: Student confidence in their spending estimate for Experiment II. In 38 instances, students gave a spending estimate with absolute confidence.

Figures 43 to 46 show statistics for the responses to the *Team Actions and Archetypes* questions (Q26–29).

59.2% of students in Experiment I said problems were handled properly by their teams, with that number increasing to 71.9% in Experiment II.

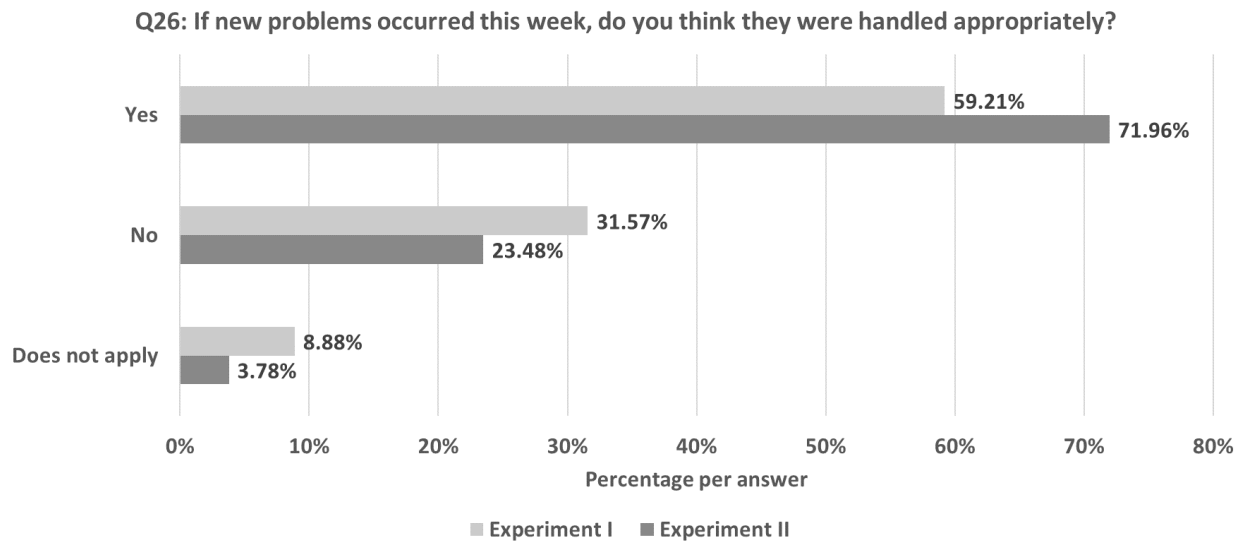


Figure 43: Statistics of the responses to Q26. The majority of students said their teams handled problems appropriately.

For Q27, there was a very close split (43.1% and 48%) during Experiment I between teams that considered new risks with new project updates and teams that did not, respectively. The results were quite different for Experiment II: 65.9% of students said they considered new potential risks in their projects. It is possible that the feedback statements, a lot of which guide students to avoid the failure causes, may have contributed to this statistic.

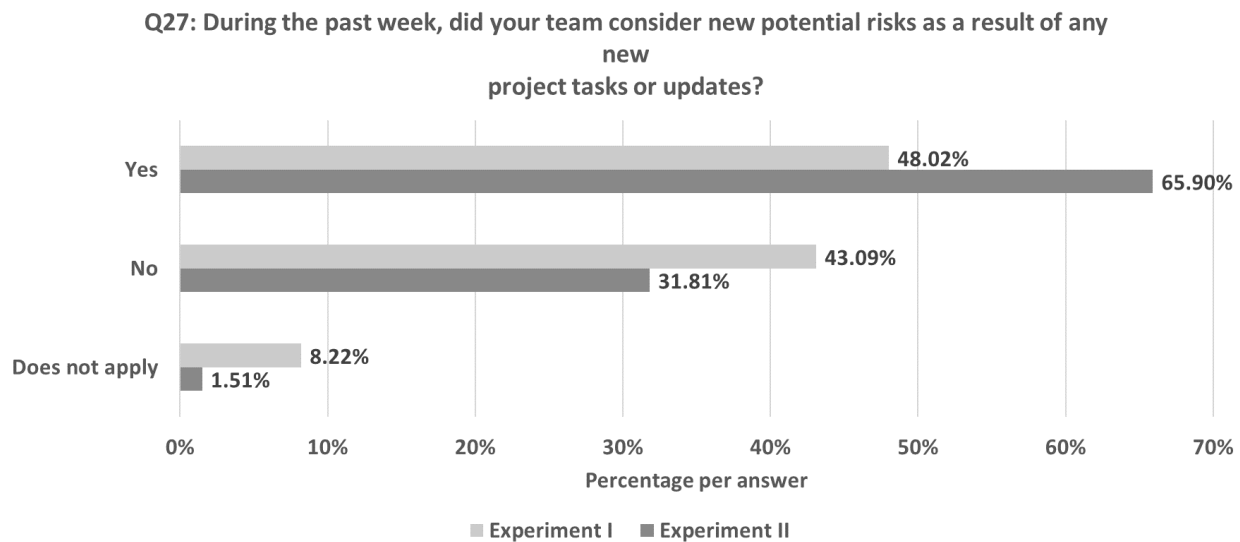


Figure 44: Statistics of the responses to Q27. Students during Experiment II did better risk management than students during Experiment I, by considering new risks to project updates. The feedback statements during Experiment II, a lot of which guide students to avoid the failure causes, may have contributed to this result.

During Experiment II, 60% of responses indicated that there were no issues with resurfacing problems, compared to 50% for Experiment I.

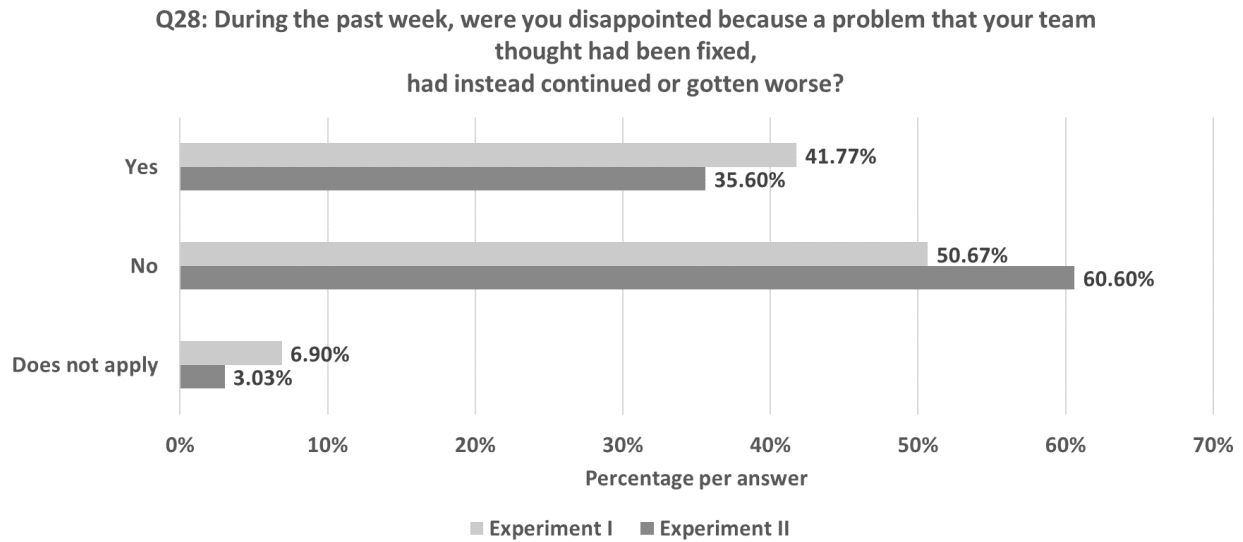


Figure 45: Statistics of the responses to Q28. At least 35% of responses indicated that students were limited by processes or rules outside their control.

Nearly 50% of students said they were frustrated by bureaucracy during experiment I, compared to 34% during Experiment II.

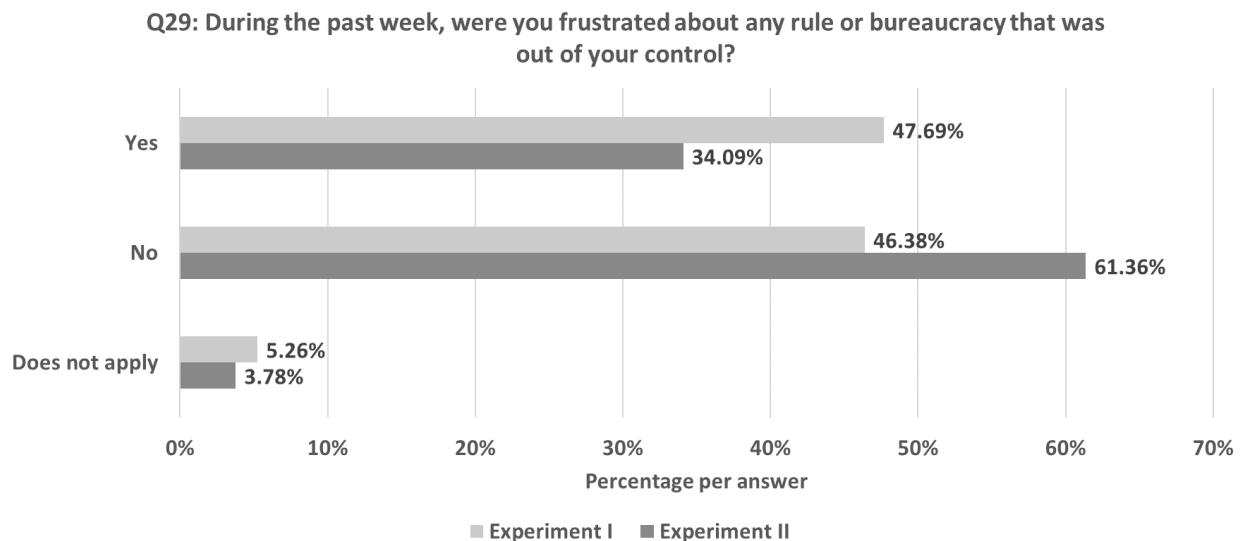


Figure 46: Statistics of the responses to Q29. During Experiment I, students were frustrated more frequently due to bureaucracy or rules.

Figures 47 to 57 show descriptive statistics for the responses to the *Indirect Signals* questions (Q30 to Q38).

40% of responses in experiment I and 45% in Experiment II show that students did not notice any change in the number of project outputs they produced. It is unsurprising that “no change” is the majority answer, given that most of the time student teams do not have vastly different number of outputs unless close to a major milestone or mishap.

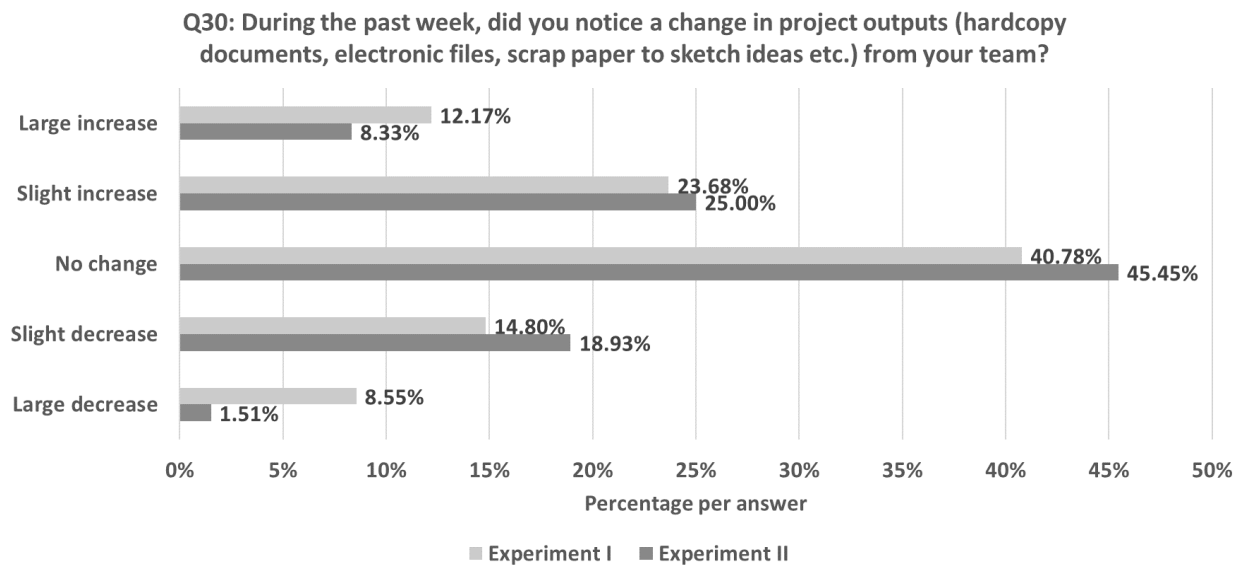


Figure 47: Statistics of the responses to Q30. Most teams likely went through all three options (decrease, no change, increase) related to the number of outputs they produced, depending on project phase. The responses show some balanced split between some increase (23% for Experiment I, 25% for Experiment II) and some decrease (14% for Experiment I and 18% for Experiment II).

Some of the indirect signals were only included in experiment I to capture some of the habits of the students in their daily lives, which may be related to how they perform during class and team meeting times. Most responses (85%) show students spent between 1 to 3 hours on social media every day, while the remaining 15% spent more than 3 hours (Figure 48). Most responses (69%) came from students who primarily ate home-prepared food, while the rest ate fast food, at Purdue's dining halls, or at restaurants (Figure 49). Most responses (55%) show students ate breakfast the day they worked on their projects and 27% did not (Figure 50).

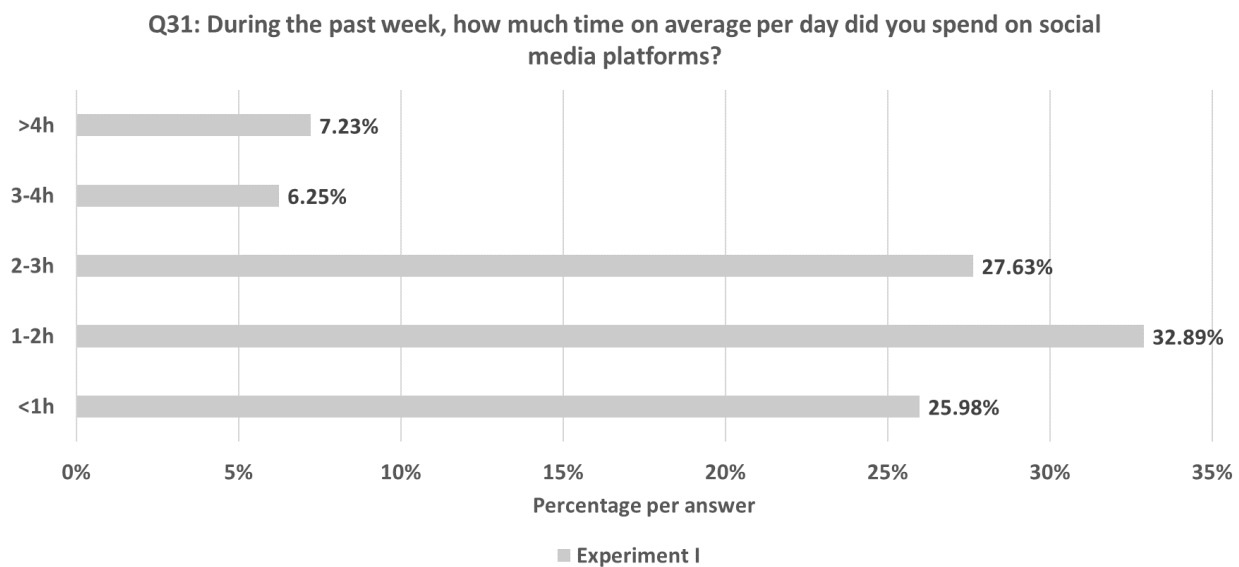


Figure 48: Statistics of the responses to Q31 during Experiment I. 85% of responses came from students who spent between 1 to 3 hours on social media, while the remaining 15% spent more than 3 hours.

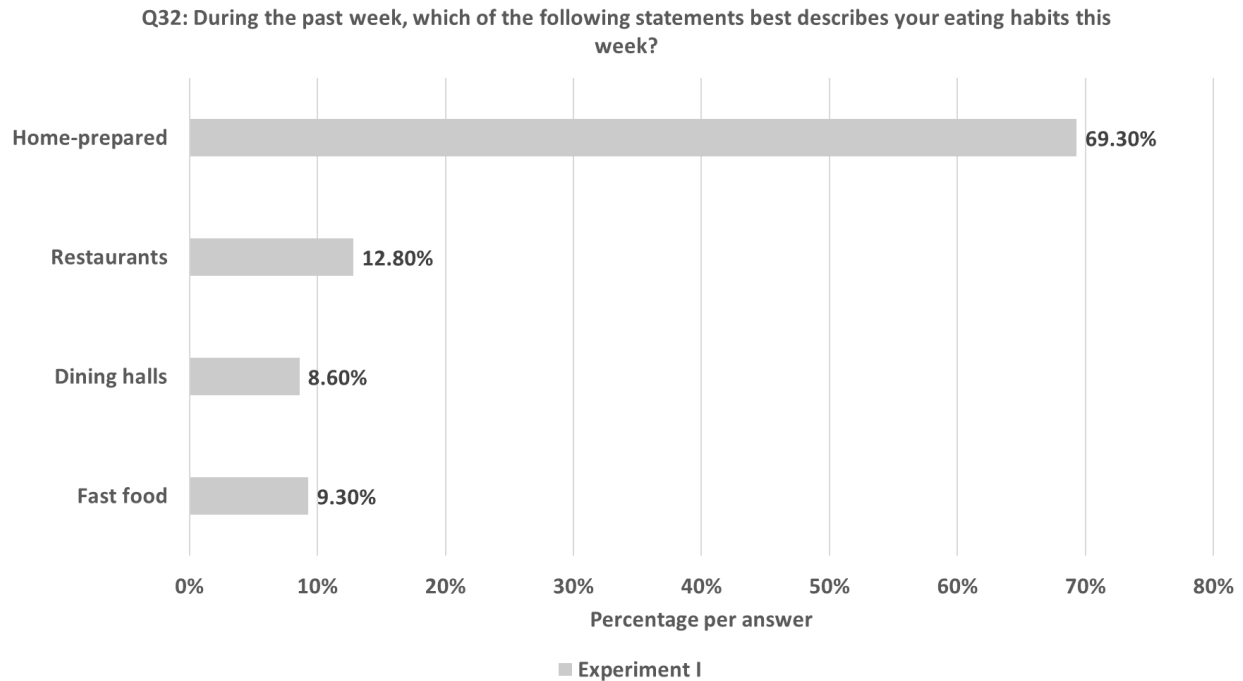


Figure 49: Statistics of the responses to Q32 during Experiment I. Based on the responses, students chose to eat fast food slightly more frequently than dining halls. The majority opted for home-prepared meals.

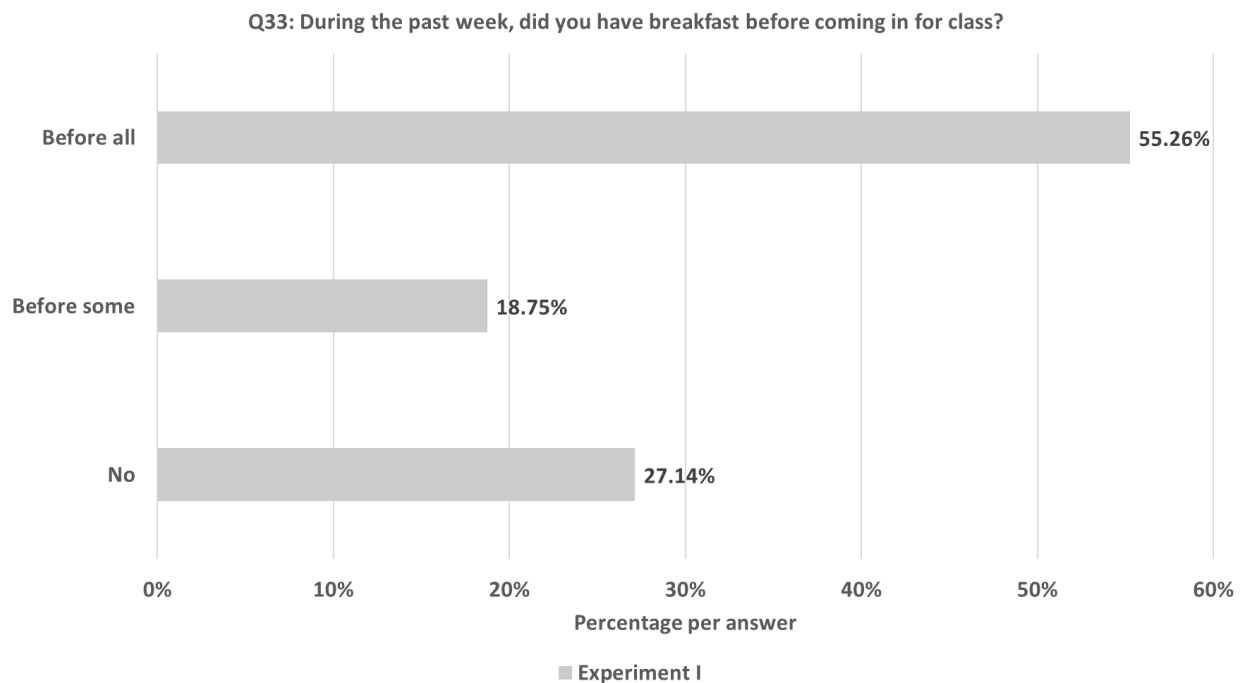


Figure 50: Statistics of the responses to Q33 during Experiment I. 72% of responses indicate that students had breakfast, at least a few times in a given week. The courses we collected data from were held in the morning, and occur 2–3 times per week.

Figures 51 and 52 show that most of the students that participated in the experiments thought about their projects 20–50% of the time, which is reasonable if one considers that they were also taking other courses.

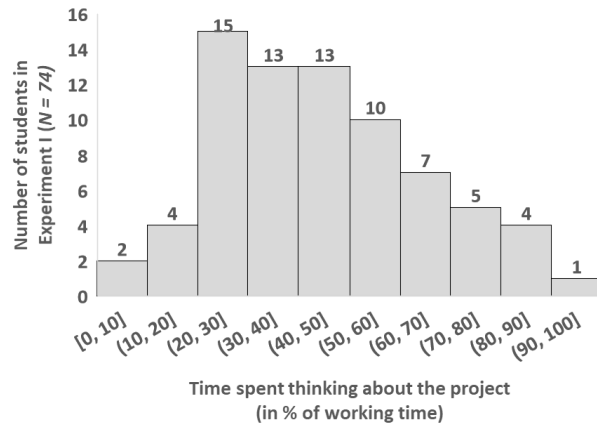


Figure 51: Most students thought about their projects between 20–50% of their working time during Experiment I, on average.

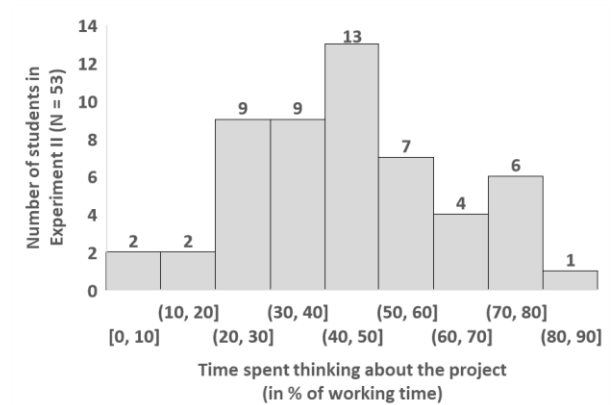


Figure 52: Similarly to Experiment I, most students thought about their projects between 20–50% of their working time during Experiment II, on average.

Figures 53 and 54 indicate that most students met up to 2 times with some members of their team outside regular class time to work on their project during Experiment I. A larger number of teams met 2 or more times during Experiment II. Design courses can be quite demanding during crunch time (e.g., preparing for equipment testing) and students need to spend more time outside class to complete their tasks.

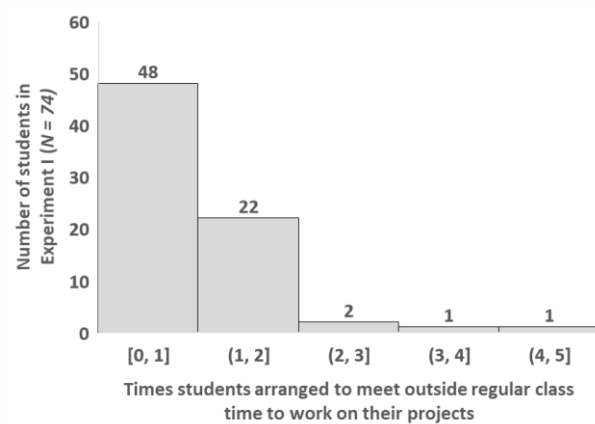


Figure 53: Most students, on average, met up to 2 times with their teams to work on their project outside regular class times during Experiment I, which is often necessary to complete demanding tasks.

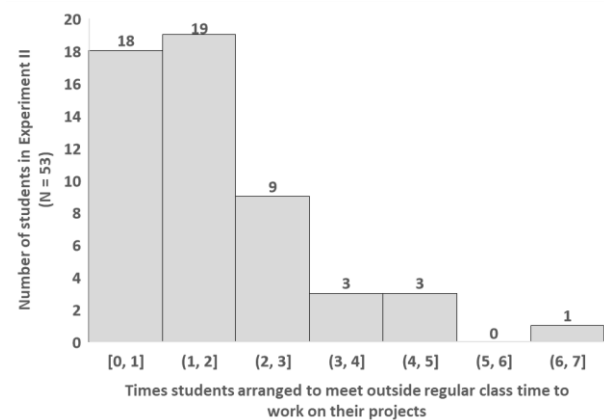


Figure 54: For Experiment II, there was larger number of teams that met outside class time for 2 or more times in a week. It is possible that students were making up for the lack of on-campus access due to COVID-19 restrictions.

Responses to Q36 show that most teams ordered between 0–5 new parts on average for most weeks during both experiments, with some occasions where a larger order was necessary.

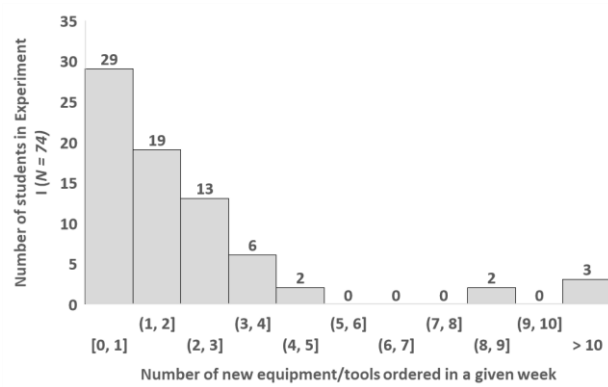


Figure 55: Most students ordered, on average, between 0–5 new parts during Experiment I, with some occasions where a larger order was necessary. These larger orders were likely sets of equipment like screws or bolts.



Figure 56: For Experiment II, the number of ordered parts per week is similar to Experiment I. Most teams ordinarily do not need to order new tools frequently.

The student responses show varying habits with respect to how often the students had physical exercise. Around 25% said they did not exercise in a particular week, while at least 74% said they exercised at least once during experiments I and II (Figure 57).

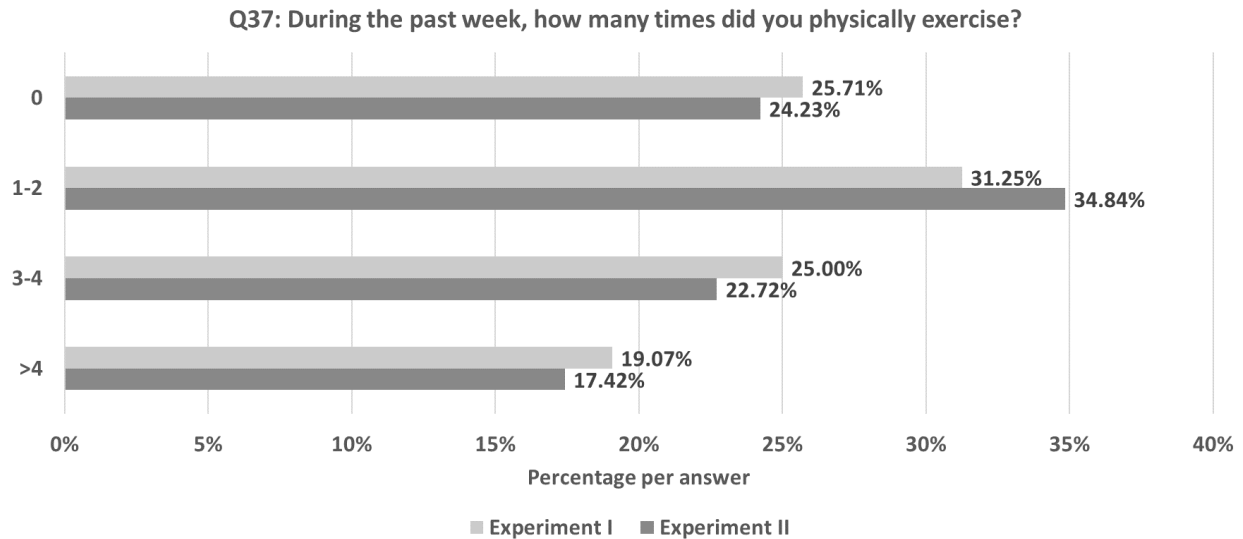


Figure 57: Majority of responses (75%) indicate that students exercised at least one time in a given week. Frequent physical exercise may correlate with improved individual performance and intellectual ability, that may influence how well students do on technical projects.

Figures 58 and 59 show descriptive statistics for the responses to the *Risk Perception* questions (Q39 to Q40).

For Q39, students were given three possible mishaps that would affect only one of the three project metrics, and they chose which one they consider the highest risk for their success at that time. 42% (Experiment I) and 37% (Experiment II) of responses show that students found a mishap in terms of technical requirements to be the highest risk. A cost mishap was considered the lowest risk of the three.

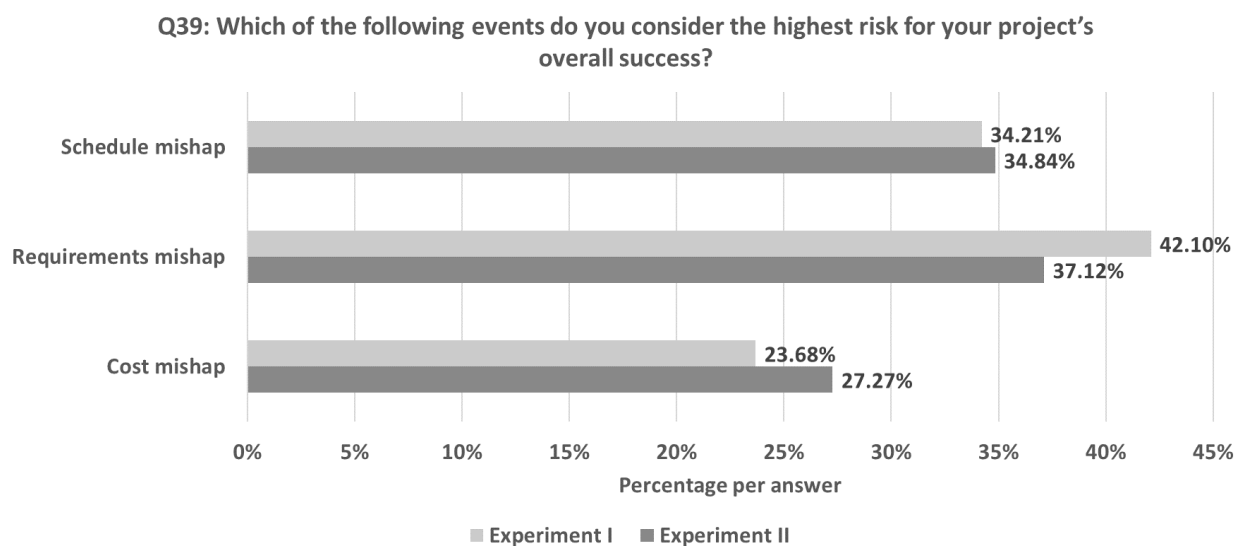


Figure 58: When ranking possible mishaps from highest to lowest risk, students considered missing their technical requirements as the highest risk during Experiment I and Experiment II, followed by a schedule mishap, and lastly a cost mishap. Budget likely mattered the least for them because they did not monitor it closely and likely did not know what the overall budget was. In contrary, they knew the timeline and requirements they had to meet, and they likely suspected they are also evaluated based on these two metrics more than how well they follow a budget.

For Q40, given the option to choose a project failure, the vast majority of responses (72% for Experiment I and 54% for Experiment II) showed preference towards budget failure, with failing to satisfy requirements the failure that the students would prefer to avoid the most.

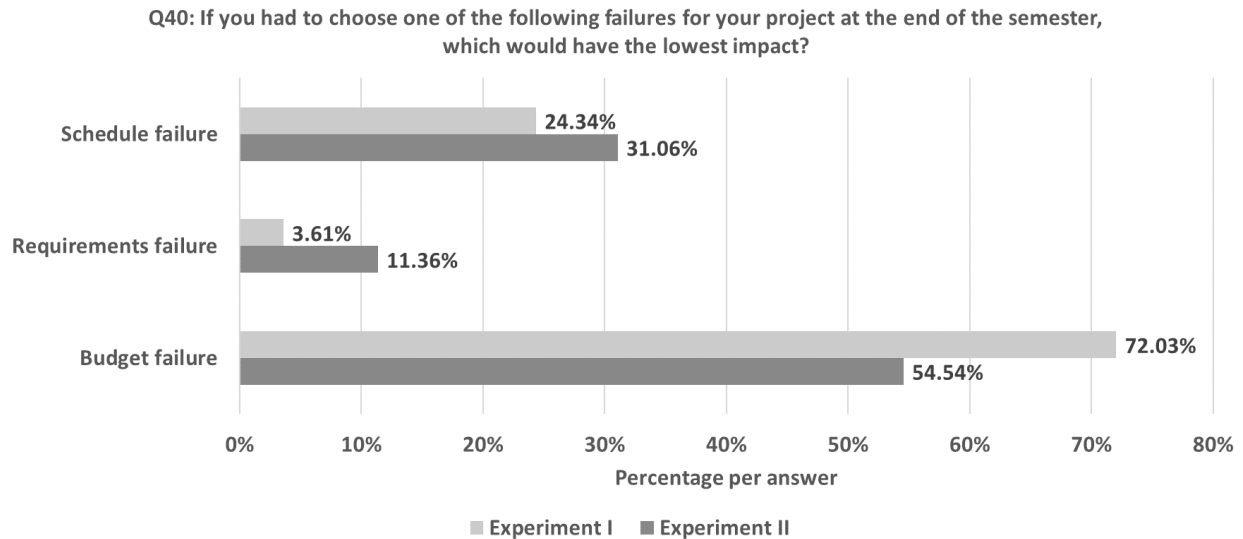


Figure 59: When asked to pick a failure they would prefer associated with the three metrics, the students gave more weight on avoiding a technical requirements failure, and most said they would rather have a budget failure.

Figures 60 to 67 show statistics for the responses to the *Individual Actions & Decisions* questions (Q41–49).

For Q41, more students said they disagreed to new ideas because of not understanding implications, indicating that students during Experiment I may have been more cautious overall, compared to Experiment II. Students not able to interact as much in the same room during Experiment II may also have contributed to the result.

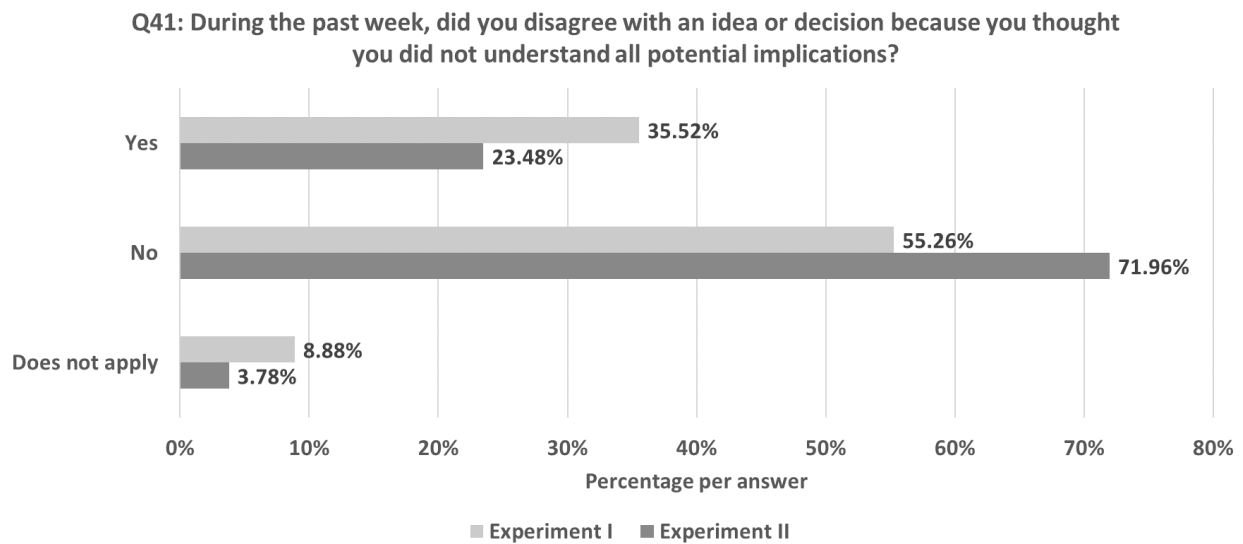


Figure 60: Statistics of the responses to Q41. 35% of responses during Experiment I indicated disagreement to new ideas when lacking understanding about potential implications, compared to 23% during Experiment II.

63% and 69% of students said they did not have arguments with their teammates (Q42) for the two experiments, respectively. We observed an almost 50%-50% split between students who singled out a decision as most important (Q43).

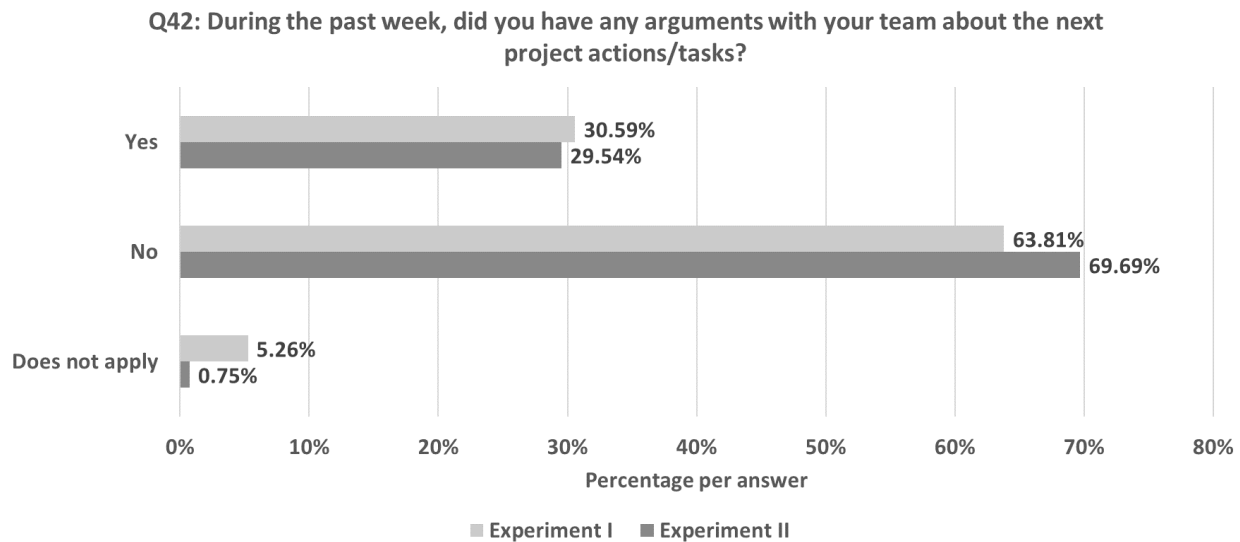


Figure 61: Statistics of the responses to Q42. Responses were similar during both experiments, majority of responses showed students not arguing during 2/3 typical weeks.

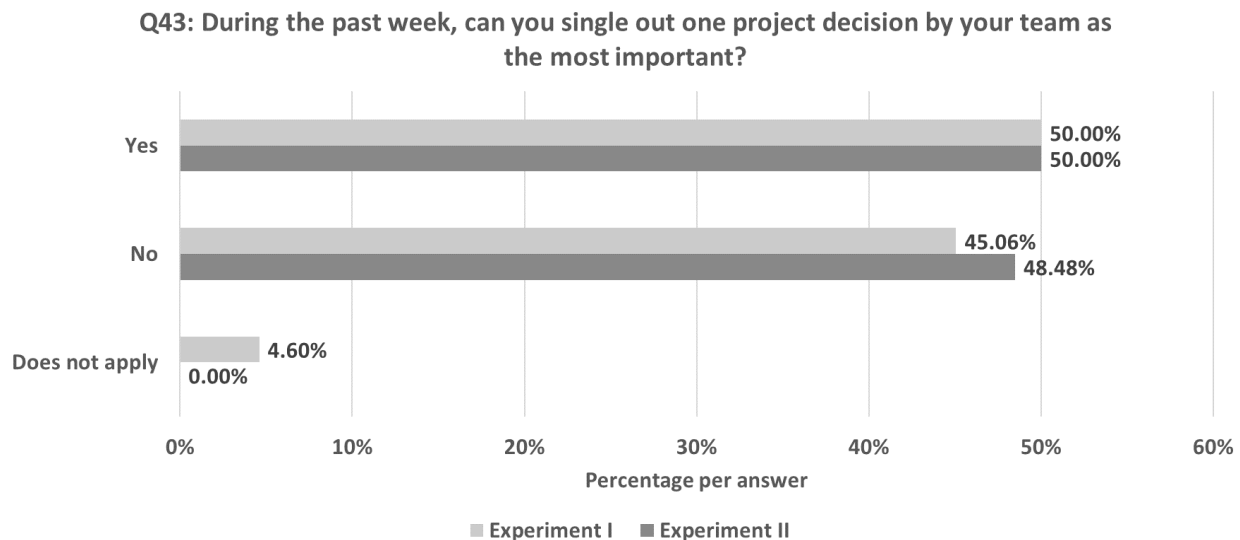


Figure 62: Statistics of the responses to Q43. Responses were similar during both experiments, with an almost perfect split between students who did and did not identify a single project decision as the most important.

Showing the value of such design courses, 57% (Experiment I) and 68% (Experiment II) of responses show students spent time thinking about what might go wrong (Q44).

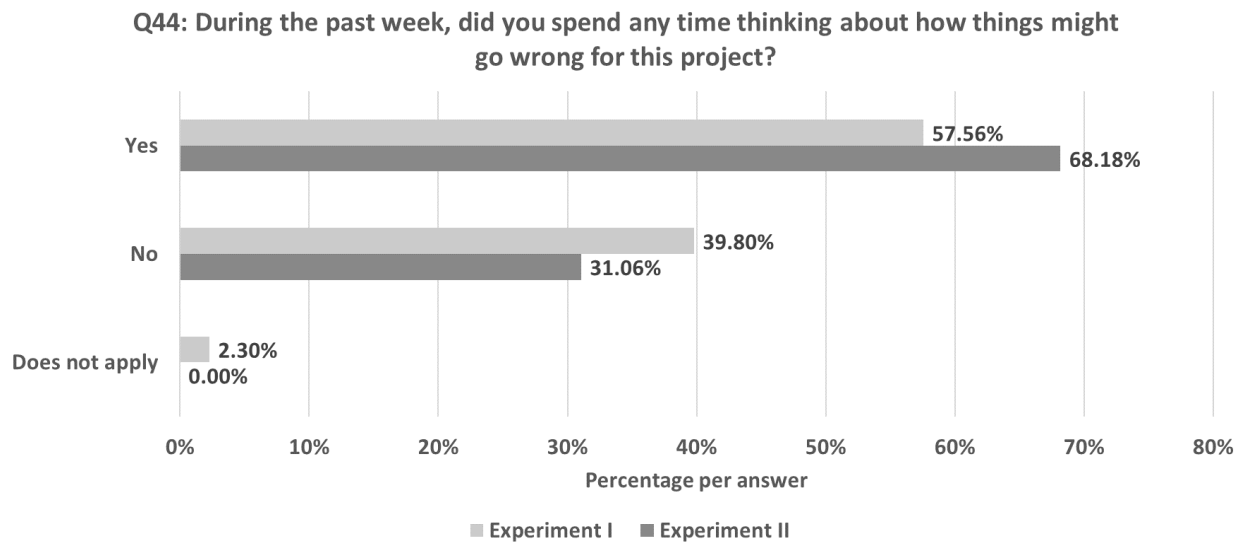


Figure 63: Statistics of the responses to Q44. 7% (Experiment I) and 68% (Experiment II) of responses show students spent time thinking about what might go wrong in their project, indicating PBL helped students develop risk management skills.

45% (Experiment I) and 68% (Experiment II) talked with other colleagues and got ideas from other teams (Q45). It may be possible that communicating with other teams was easier during Experiment II because of ease of scheduling virtual meetings between students.

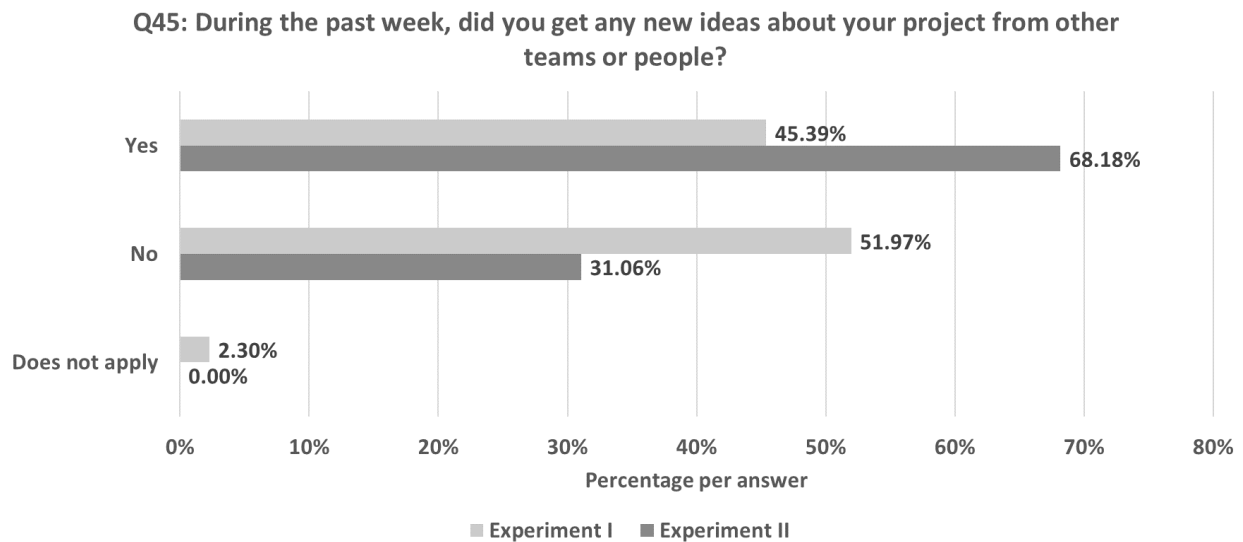


Figure 64: Statistics of the responses to Q45. 45% (Experiment I) and 68% (Experiment II) talked with other colleagues and got ideas from other teams.

51% (Experiment I) and 49% (Experiment II) of responses showed that students learned something that surprised them (Q46).

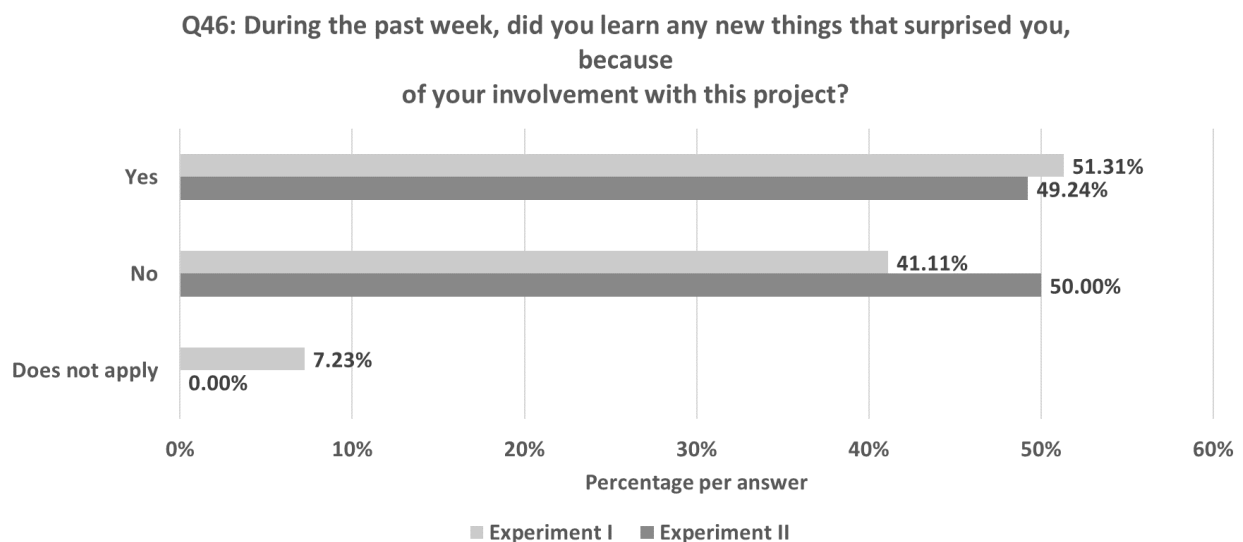


Figure 65: Statistics of the responses to Q46. Almost for half their time involved in the projects, students learned something new.

65% of responses during Experiment II (compared to 58% during Experiment I) indicated that students did not spend time discussing what they thought as trivial matters to the project. The results show mostly equal time efficiency between the two experiments.

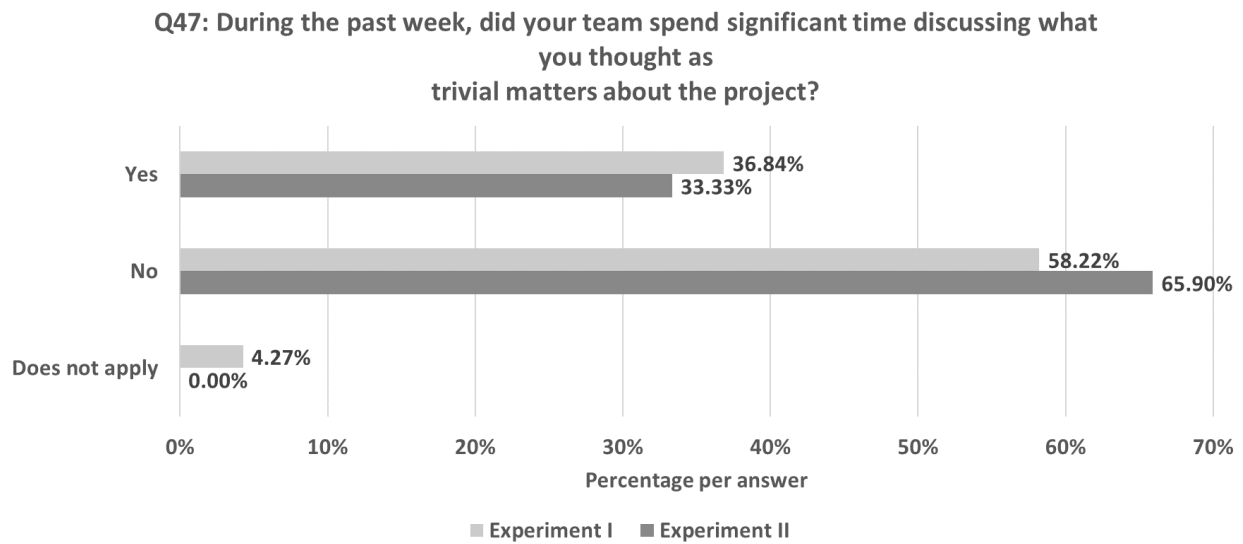


Figure 66: Statistics of the responses to Q47. The results show mostly equal time efficiency between the two experiments.

More than 67% said they thought through all solutions before proceeding with an action (Q48).

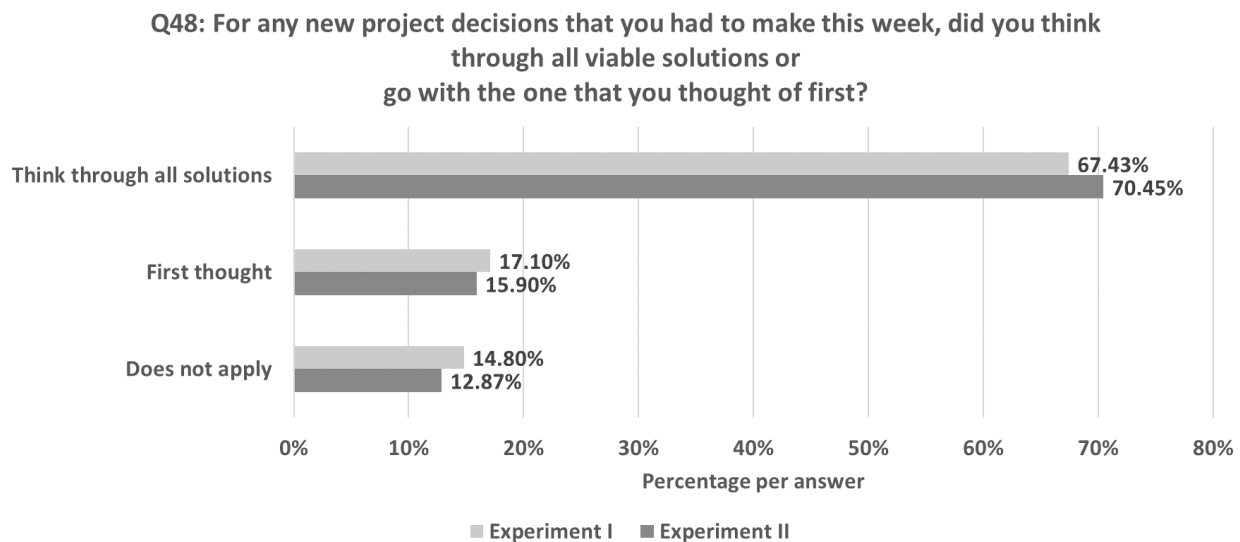


Figure 67: More than 2/3 of the responses come from students who said they thought through all solutions before making a decision during both experiments, with some indicating the question did not apply for a particular week, suggesting that there was no project decision for them to make.

Overall, many questions have similar response patterns between the two experiments with a few exceptions. When questioned on meaningful progress and interaction with their team, students from Experiment II appeared to do better than the responses from Experiment I, which may indicate that feedback (targeted or not) helps students improve their coordination when working together and as a result make meaningful progress. During Experiment II, students did not know as much about their team members activities as they did during Experiment I. Hybrid learning and safety measures due to the COVID-19 pandemic may be responsible for that, given students have to follow capacity limits in their workspaces so they cannot work together as much. During Experiment I, the students thought that the objectives for the projects were a lot more clearly defined with less room of ambiguity. During Experiment II, the courses and instructors were different than those for Experiment I, which may be a contributing factor to the clarity of the objectives. In Experiment II, 12% of responses showed that students would definitely abandon the project (compared to only 5% in Experiment I) if they had to start with a completely new team, which shows less commitment from them.

7. CONCLUSIONS AND FUTURE WORK

The research work discussed in this dissertation is an effort to improve multiple facets of project-based learning (PBL) courses, with the goal of better preparing engineering students to learn to deal with failures before they join the workforce. Industry projects going over budget, missing schedule milestones, or falling short of meeting requirements is a frequent phenomenon, and making better future engineers may be one of the most effective means we have to remedy such trends. To improve PBL courses I followed two main pathways: 1) find areas that PBL could improve at preparing students for failures and suggest improvements to instructors, and 2) understand student behaviors that lead to PBL project failures and provide helpful feedback so students can improve.

To investigate the actions and behaviors that lead to failure events in student projects, I introduced “crowd signals”, which is crowdsourced information collected directly from the students that are part of a project team. To arrive at a successful set of crowd signals, I surveyed literature looking for factors that affect team, project, and individual performance and developed 49 questions to collect the crowd signals. Each factor then led to one or more student questions that applied specifically to the specialized context of student projects. The crowd signal inputs together with project performance information from instructors were necessary to complete the research work discussed here. To collect the necessary data and analysis, I completed two experiments at Purdue University including student teams from PBL courses: the first included 28 student teams from two courses and the second included 14 student teams from three courses.

The first part of this dissertation (Chapter 3) identified potential areas that PBL could improve upon, and suggested specific recommendations to instructors that want to enhance the educational value of their PBL courses. I compared industry failure cause occurrence rates with similar rates from student teams. Failure causes refer to events that frequently caused budget, schedule, or requirements failures in industry, and were identified from previous research. My analysis showed that of the 10 failure causes I measured, *failed to consider a design aspect* statistically appears less frequently in the student projects compared to industry. I then built logistic regression models to find the correlations between the crowd signals and the occurrence of the failure cause *failed to*

consider a design aspect and used that information as a guide to suggest improvements to instructors. I provided four suggestions to instructors that want to provide their students with more opportunities to learn from the failure cause:

1. The instructor could arrange the course so that specific tasks happen in settings where students do not interact with each other as much.
2. The instructor could introduce reporting rules and processes that resemble industry standards.
3. The instructor could implement realistic equipment/tool usage and expense constraints to the student teams.
4. The instructor could put more emphasis on the requirements of the project, the importance of them, and how they clearly relate to project success.

My goal for the recommendations is to help students get more out of their education by getting more opportunities to experience failure safely before they join industry projects. To some instructors, depending on how they evaluate their courses, these changes may appear as negative. Despite the seemingly negative notion of failure, these changes may increase the effectiveness and educational value of PBL. Instructors can integrate such changes in a controlled manner to retain a fair grading scheme for their course (e.g., consistently applying changes across all projects or evaluating students based on effort or evidence of learning from the failure and improvement, rather than a project metric).

The second part of this dissertation (Chapter 4) stems from an effort to understand which student behaviors lead to which types of project failures in PBL courses, and to gauge whether the crowd signals are good predictors of future project performance. I developed logistic regression models that predict the occurrence of future budget, schedule, or requirements failures, using crowd signals and other information as inputs, and evaluated those models to get an insight on which student actions are likely to lead to project failures. The models predict, on average, $73.11 \pm 6.92\%$ of budget failures, $75.27\% \pm 9.21\%$ of schedule failures, and $76.71 \pm 6.90\%$ of technical requirements failures after reducing the inputs via a hybrid approach of stepwise elimination and best subsets.

The initial model that predicts budget failure indicated that when students perceive they have increased freedom on what to do with the project and do not have problems continue or become

worse, the likelihood of a cost failure reduces. In contrast, budget failure likelihood increases when students perceive a schedule failure as higher risk compared to a cost failure, when they disagree because of lack of understanding decision implications, when they single out a decision as most important, and when having a budget failure the previous week.

The initial model that predicts schedule failure indicated that when students are sharing about their lives, think they are spending more funds than they should, are turning down activities that they consider fun, and understand all potential implications of an action, a schedule failure is less likely. On the contrary, with increasing student confidence in their success without oversight, coming up with or agreeing to more project ideas, thinking they are satisfying requirements as planned, having previous problems resurface due to poor previous solutions, spending more than 4 hours on social media per day, perceiving schedule as the highest risk for the project, not learning any new things, and having a schedule failure the previous week, all increase the likelihood of a schedule failure.

The initial model that predicts failure regarding the technical requirements indicated that not exercising at all during the week, discussing trivial matters during the project and being increasingly confident in one's answers to the questions reduce the likelihood of a failure. In contrast, when students are increasingly unable to focus on the project, introduce new ideas to the project, skip or postpone required tasks, think they are spending more than they should, report their cost estimate with high confidence, and having a requirements failure the previous week, the likelihood of a future requirements failure increases.

Lastly, the last contribution (Chapter 5) focused on improving student behaviors to potentially improve PBL project performance. To accomplish this goal, I generated 35 feedback statements, guided from the correlations between failure measures and the crowd signals. The student teams were split in two treatment groups: teams that received targeted feedback (i.e., feedback that aimed to address the failure causes that the specific team is most prone to) and teams that received non-targeted feedback (i.e., feedback that is positive but does not necessarily address the failure causes the specific team is most prone to). I used the second experiment to evaluate whether the targeted feedback helps reduce the occurrences of project failures in the student teams.

Through my analysis, I found that my targeted feedback does not reduce the failure occurrences in terms of any metrics. The quantitative and qualitative results indicated that student teams who received the targeted feedback statements said they were more likely to change their behavior, but the project success rates did not improve overall, suggesting that the students either did not make any changes or whatever changes they made were ineffective, perhaps because they did not know how to change their behavior appropriately. Students also said the targeted feedback was more helpful than the non-targeted feedback, but the project success rates indicate that they were only able to improve in terms of the schedule metric. Lastly, more students in the targeted feedback group said that the feedback statements would have a positive impact on their projects, but end-of-semester success rates did not confirm their responses. It is possible that the feedback does contribute positively to the teams, but is not enough to have an impact on project success.

7.1 Limitations

The research presented here comes with some limitations. A major source of these limitations relates to the truthfulness in the responses of the respondents (students and instructors) as well as their capability to provide good responses to the questions. I assumed instructors to be in a position where they are able to detect occurrences of failures and failure causes occurring in the teams. For the courses included in my experiments, the instructors are heavily involved in what happens with the teams and monitor them multiple times a week. However, I had no way of enforcing the frequency or depth of such interactions. Also, the students who responded to my surveys indirectly impact the capability of the models and feedback evaluation, which was a major part of my research. I had no way of monitoring whether a student would willingly respond untruthfully, although I do not have a valid reason to expect that.

Another source of limitations for the research comes due to the experiments including instructors and courses from the same department, potentially resulting in skewed conclusions. It is possible that if the study were to be repeated in other courses or departments, the results would differ in some way. My intention with the conclusions of this research work is not to generalize into other applications or settings, but rather to simply report my findings.

Lastly, the quantitative and qualitative results of the feedback evaluation, particularly due to the small number of observed projects, are dependent to the types of projects and ability of the student teams, as well as the effects of randomly assigning teams in the two treatment groups.

7.2 Suggestions for Future Research

I made an initial attempt at understanding the failure mechanisms in project-based learning and suggesting improvements to instructors (as PBL improvement recommendations) and to students (as feedback). The current section discusses improvements and viable areas of future research.

7.2.1 Extension to student projects in multiple disciplines

For both experiments, I collected data from teams that were working on aerospace-themed projects, because of my experience and involvement with the department in which I was a graduate student. One suggestion would be to expand the data collection to student projects in other engineering disciplines at first, and then to disciplines outside engineering. Mechanical, electrical, and industrial engineering disciplines would be viable since the projects do resemble those offered in aerospace engineering. For this first step, it would be of value to see how similar, or different, the project failure mechanisms are for student teams among engineering courses. As for the second step, computer science or other projects that do not involve building equipment could be a viable option. In that case, some of the research would need to change (e.g., removing hardware-related feedback). The goal with this second step, would be to identify key differences for failure in projects that are equally complex as engineering projects, but without hardware-related activities such as manufacturing.

7.2.2 Automated feedback process

With decreasing instructor time and increasing student enrollment, my approach of targeted feedback can be of benefit to instructors and students. However, the feedback generation process is currently semi-automated, as I updated the surveys and ran the prediction models every week and for every team. Removing any kind of human input by writing the appropriate algorithms to receive the necessary data, run the models, and update the feedback on the survey would

completely automate the process. An entirely automated process could be an attractive option for instructors who want to help their students with supplementary feedback.

7.2.3 Improving predictive models

For the predictive models, I used logistic regression, which served the purpose of allowing interpretation of the correlations between crowd signals and failure measures, which was a necessary step for the research. There are many other classification methods, including neural network architectures, that potentially have much stronger capability at capturing the correlations of the problem and therefore make stronger predictions. With the knowledge that there is some merit to crowdsourced information and its usefulness to failure prediction, I recommend that future efforts focus on classification methods that can produce much more reliable predictions.

7.2.4 Development of an integrated app environment

A significant portion of the research involved simple survey tools and researcher intervention (e.g., for updating the feedback statements for the student teams). A future project would be to migrate all these functions into an app that can be accessed as a web environment, which would allow for more interactive features for the predictions and the feedback.

7.2.5 Industry setting

Another viable research question is to determine whether the results from the student teams are similar to professional engineering teams and to assess whether the risk assessment approach discussed in this dissertation can help organizations deal better with and reduce failures. Industry partners could extract value by leveraging the predictive capability of the prototype and generating feedback to decision makers, alerting them of upcoming failures, and suggesting corrective actions that are tailored to the organization. The approach outlined in this dissertation provides the added benefit of giving insight into the mechanisms of risk at the specific organization (since the predictive models would be trained with internal employee data).

In the industry application, it would also be relevant to add additional inputs that come from traditional risk management tools that are used in industry (e.g., financial software or product

quality tracking software), to the predictive models. An industry setting would also allow for more, and more consistent data, if combined with a user-friendly app setting (Figure 68).

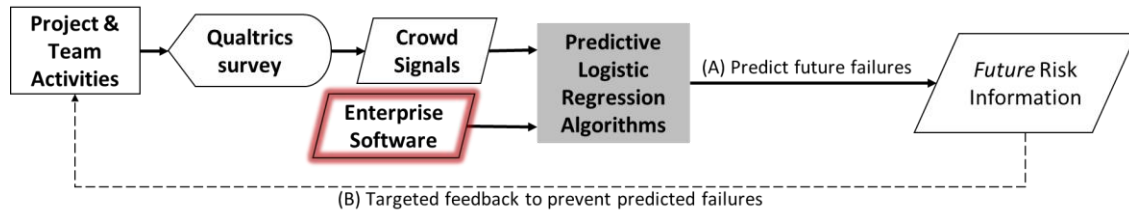


Figure 68: Industry prototype for project failure prediction and prevention.

7.2.6 Introducing quantitative metrics of failure

One potential improvement is to also consider specific quantitative metrics in addition to the binary failure outcomes that I used in this work. Industry projects are over budget by a specific percentage, behind schedule by a certain number of months, and are missing a specific number of requirements. Therefore, it would be helpful to consider by how much more or less a particular crowd signal contributes towards a specific failure compared to the present status, to be able to predict, for example how many more months is a project going to be behind schedule given the performance of the team for the past few weeks.

7.2.7 Causal mechanisms between crowd signals and failure measures

The work presented here focused on the correlations between crowd signals (that aim to measure specific factors) and failure metrics. It would be worthwhile for future research to focus on whether causation between these underlying factors and failures is also true, as knowledge about such relationships can directly impact project management decisions.

For example, I included agreeableness as a factor that may impact team performance. I formed one possible question to measure agreeableness (Q19), which asked whether team members share information about each other's lives. The schedule model showed that with increasing frequency of team members sharing more about each other's lives, the likelihood of schedule failures reduces. The result indicates that schedule failure and the responses to Q19 are dependent, but do not necessarily have a causal relationship. In a hypothetical student team where team members are not

spending much time together discussing various topics about their lives, we do not know with certainty that the project will be late. Also, if I used a different question to measure agreeableness, I may not have arrived at the same result.

Therefore, there are two main areas of valid research in the topic of causal mechanisms between crowd signals and failure measures:

1. Find the types of questions that measure the factors I considered well (*get the right crowd signal to measure the factor*) and
2. Conduct targeted experiments to investigate whether these specific factors cause failure (*does the factor cause failure*).

APPENDIX A. CS-FC (“CROWD SIGNAL—FAILURE CAUSE”) CORRELATIONS

Appendix A includes the logistic regression correlation coefficients that I used in the CS-FC matrix referred to in Section 5.4. The coefficient values were the guides to developing the feedback statements for Experiment II. I built the models presented here using the exact same approach as described for FC1: *Failed to consider a design aspect* in Section 3.2, and correspond to the remaining failure causes FC2 to FC10.

FC3: *Failed to form a contingency plan* does not have a model as it did not occur enough times for the model to converge (occurrence ratio in model training dataset = 0.107) and therefore, I excluded it from the feedback generation process. I also excluded FC9: *Violated procedures* (occurrence ratio in model training dataset = 0.064) for the same reason.

Table 35: Mixed-effects logistic regression model coefficients for FC2: Used inadequate justification.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
<i>a</i>	-1.845 (2.889)	<i>b</i> ₃₀ (Q26 = No)	0.220 (0.986)	<i>b</i>₆₀(Q41 = Yes)	2.162 (1.066) *
<i>b</i> ₁ (Q1)	-0.328 (0.226)	<i>b</i> ₃₁ (Q26 = Yes)	-0.690 (0.918)	<i>b</i> ₆₁ (Q42 = No)	-1.642 (1.118)
<i>b</i>₂(Q2)	0.382 (0.193)*	<i>b</i> ₃₂ (Q27=No)	-0.146 (1.100)	<i>b</i> ₆₂ (Q42 = Yes)	-1.443 (1.150)
<i>b</i> ₃ (Q3)	0.423 (0.223)^	<i>b</i> ₃₃ (Q27=Yes)	0.186 (1.022)	<i>b</i> ₆₃ (Q43 = Yes)	0.297 (1.455)
<i>b</i> ₄ (Q4)	0.026 (0.226)	<i>b</i> ₃₄ (Q28=No)	1.849 (1.241)	<i>b</i> ₆₄ (Q44 = Yes)	1.737 (1.795)
<i>b</i> ₅ (Q5)	0.090 (0.223)	<i>b</i> ₃₅ (Q28=Yes)	0.522 (1.272)	<i>b</i>₆₅(Q45 = Yes)	-4.503 (2.274) *
<i>b</i> ₆ (Q6)	0.418 (0.245)^	<i>b</i> ₃₆ (Q29=No)	0.601 (1.498)	<i>b</i> ₆₆ (Q46 = No)	0.483 (1.214)
<i>b</i> ₇ (Q7)	0.257 (0.230)	<i>b</i> ₃₇ (Q29=Yes)	0.472 (1.584)	<i>b</i> ₆₇ (Q46 = Yes)	0.178 (1.214)
<i>b</i> ₈ (Q8)	-0.088 (0.237)	<i>b</i> ₃₈ (Q30)	-0.074 (0.210)	<i>b</i>₆₈(Q47 = No)	3.530 (1.630) *
<i>b</i> ₉ (Q9 = Low)	0.016 (1.758)	<i>b</i> ₃₉ (Q31 =2-3h)	-0.744 (0.553)	<i>b</i> ₆₉ (Q47 = Yes)	3.142 (1.692) ^
<i>b</i> ₁₀ (Q9 = Moderate)	0.703 (0.437)	<i>b</i> ₄₀ (Q31 = 3-4h)	0.217 (0.862)	<i>b</i> ₇₀ (Q48 = First thought)	-0.57 (0.793)
<i>b</i> ₁₁ (Q10)	0.337 (0.213)	<i>b</i>₄₁(Q31 = <1h)	-1.206 (0.582)*	<i>b</i> ₇₁ (Q48 = Think through)	-0.089 (0.649)
<i>b</i> ₁₂ (Q11)	-0.239 (0.228)	<i>b</i> ₄₂ (Q31 = >4h)	-1.544 (0.934)^	<i>b</i> ₇₂ (Q49)	0.515 (0.264) ^
<i>b</i>₁₃(Q12)	0.598 (0.245)*	<i>b</i> ₄₃ (Q32 = Dining hall)	-0.567 (0.924)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.2104 Random effects $c_i \sim N(0, 0.513^2)$ Occurrence ratio in data: 0.232</p>	
<i>b</i> ₁₄ (Q13)	-0.242 (0.214)	<i>b</i> ₄₄ (Q32 = Restaurants)	-1.711 (0.97)^		
<i>b</i> ₁₅ (Q14)	0.074 (0.227)	<i>b</i> ₄₅ (Q32 = Home)	-1.378 (0.704)^		
<i>b</i> ₁₆ (Q15)	-0.349 (0.253)	<i>b</i> ₄₆ (Q33=No)	0.471 (0.563)		
<i>b</i> ₁₇ (Q16)	-0.258 (0.215)	<i>b</i>₄₇(Q33=Some)	1.500 (0.570) **		
<i>b</i> ₁₈ (Q17)	-0.393 (0.27)	<i>b</i> ₄₈ (Q34)	-0.385 (0.247)		
<i>b</i> ₁₉ (Q18)	0.302 (0.246)	<i>b</i>₄₉(Q35)	0.813 (0.259) **		
<i>b</i> ₂₀ (Q19)	-0.415 (0.221)^	<i>b</i> ₅₀ (Q36)	-0.053 (0.239)		
<i>b</i> ₂₁ (Q20 = Over budget)	0.273 (0.693)	<i>b</i> ₅₁ (Q37 = >3-4)	0.109 (0.607)		
<i>b</i> ₂₂ (Q20 = Under budget)	-0.428 (0.534)	<i>b</i> ₅₂ (Q37 = >4)	0.868 (0.640)		
<i>b</i>₂₃(Q21 = Behind sched.)	-2.206 (0.836)**	<i>b</i> ₅₃ (Q37 = None)	0.592 (0.608)		
<i>b</i>₂₄(Q21 = On sched.)	-1.949 (0.733)**	<i>b</i> ₅₄ (Q38)	-0.352 (0.243)		
<i>b</i> ₂₅ (Q22 = More reqs)	1.261 (0.84)	<i>b</i> ₅₅ (Q39=Reqs)	-1.164 (0.617)^		
<i>b</i> ₂₆ (Q22=reqs as planned)	0.340 (0.662)	<i>b</i> ₅₆ (Q39=Sched)	-0.439 (0.53)		
<i>b</i> ₂₇ (Q23)	0.112 (0.253)	<i>b</i>₅₇(Q40=Reqs)	2.820 (1.181) *		
<i>b</i> ₂₈ (Q24)	-0.185 (0.281)	<i>b</i> ₅₈ (Q40=Sched)	0.643 (0.508)		
<i>b</i> ₂₉ (Q25)	0.350 (0.276)	<i>b</i> ₅₉ (Q41 = No)	1.216 (1.026)		

Table 36: Mixed-effects logistic regression model coefficients for FC4: Lacked Experience

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	3.698 (2.385)	b_{30} (Q26 = No)	-1.012 (1.089)	b_{60} (Q41 = Yes)	0.012 (1.096)
b_1 (Q1)	0.048 (0.248)	b_{31} (Q26 = Yes)	-1.181 (0.925)	b_{61} (Q42 = No)	-0.689 (1.122)
b_2 (Q2)	-0.424 (0.311)	b_{32} (Q27=No)	0.03 (1.032)	b_{62} (Q42 = Yes)	-1.506 (1.223)
b_3 (Q3)	0.076 (0.238)	b_{33} (Q27=Yes)	-0.217 (0.996)	b_{63} (Q43 = Yes)	0.146 (1.488)
b_4 (Q4)	-0.164 (0.245)	b_{34} (Q28=No)	1.661 (1.397)	b_{64} (Q44 = Yes)	0.199 (1.598)
b_5 (Q5)	0.137 (0.223)	b_{35} (Q28=Yes)	1.194 (1.49)	b_{65} (Q45 = Yes)	0.683 (2.29)
b_6 (Q6)	0.23 (0.223)	b_{36} (Q29=No)	-1.201 (1.423)	b_{66}(Q46 = No)	-2.91 (0.994)**
b_7 (Q7)	-0.267 (0.229)	b_{37} (Q29=Yes)	-1.174 (1.452)	b_{67}(Q46 = Yes)	-2.439 (0.95)*
b_8 (Q8)	0.097 (0.229)	b_{38} (Q30)	-0.148 (0.216)	b_{68} (Q47 = No)	-0.795 (1.399)
b_9 (Q9 = Low)	0.889 (1.481)	b_{39} (Q31 =2-3h)	-0.828 (0.597)	b_{69} (Q47 = Yes)	-1.586 (1.473)
b_{10} (Q9 = Moderate)	0.127 (0.457)	b_{40} (Q31 = 3-4h)	1.439 (0.924)	b_{70} (Q48 = First thought)	0.57 (0.742)
b_{11} (Q10)	0.224 (0.214)	b_{41} (Q31 = <1h)	-0.166 (0.562)	b_{71} (Q48 = Think through)	-0.074 (0.659)
b_{12} (Q11)	-0.019 (0.221)	b_{42} (Q31 = >4h)	-0.188 (0.889)	b_{72} (Q49)	0.257 (0.233)
b_{13} (Q12)	0.087 (0.238)	b_{43} (Q32 = Dining hall)	-0.234 (1.01)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.245 Random effects $c_i \sim N(0, 0.468^2)$ Occurrence ratio in data: 0.217</p>	
b_{14} (Q13)	0.081 (0.233)	b_{44} (Q32 = Restaurants)	-0.008 (0.95)		
b_{15} (Q14)	-0.436 (0.233)^	b_{45} (Q32 = Home)	-0.144 (0.744)		
b_{16} (Q15)	-0.128 (0.252)	b_{46} (Q33=No)	-0.464 (0.585)		
b_{17} (Q16)	0.08 (0.219)	b_{47} (Q33=Some)	1.001 (0.6)^		
b_{18} (Q17)	0.07 (0.239)	b_{48} (Q34)	0.132 (0.236)		
b_{19} (Q18)	-0.242 (0.226)	b_{49}(Q35)	0.642 (0.234)*		
b_{20} (Q19)	-0.225 (0.217)	b_{50} (Q36)	0.126 (0.238)		
b_{21} (Q20 = Over budget)	-1.089 (0.817)	b_{51} (Q37 = >3-4)	0.572 (0.595)		
b_{22} (Q20 = Under budget)	-0.401 (0.557)	b_{52} (Q37 = >4)	-1.332 (0.829)		
b_{23} (Q21 = Behind sched.)	-0.619 (0.831)	b_{53} (Q37 = None)	0.745 (0.596)		
b_{24} (Q21 = On sched.)	-0.229 (0.772)	b_{54} (Q38)	0.046 (0.238)		
b_{25} (Q22 = More reqs)	-0.845 (0.943)	b_{55} (Q39=Reqs)	-1.099 (0.644)^		
b_{26} (Q22=reqs as planned)	-0.429 (0.616)	b_{56} (Q39=Sched)	0.235 (0.546)		
b_{27} (Q23)	-0.011 (0.254)	b_{57} (Q40=Reqs)	0.312 (1.083)		
b_{28}(Q24)	0.657 (0.277)*	b_{58} (Q40=Sched)	-0.252 (0.558)		
b_{29} (Q25)	-0.448 (0.285)	b_{59} (Q41 = No)	-0.238 (1.055)		

Table 37: Mixed-effects logistic regression model coefficients for FC5: Kept poor records.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
<i>a</i>	2.787 (7.117)	<i>b</i> ₃₀ (Q26 = No)	0.232 (2.306)	<i>b</i>₆₀(Q41 = Yes)	3.942 (1.459) **
<i>b</i> ₁ (Q1)	0.389 (0.456)	<i>b</i> ₃₁ (Q26 = Yes)	-3.266 (2.174)	<i>b</i> ₆₁ (Q42 = No)	-5.433 (2.856) ^
<i>b</i> ₂ (Q2)	-0.742 (0.651)	<i>b</i> ₃₂ (Q27=No)	-0.139 (2.228)	<i>b</i> ₆₂ (Q42 = Yes)	-2.530 (2.406)
<i>b</i> ₃ (Q3)	0.358 (0.428)	<i>b</i> ₃₃ (Q27=Yes)	2.174 (2.262)	<i>b</i> ₆₃ (Q43 = Yes)	-4.183 (2.621)
<i>b</i> ₄ (Q4)	0.920 (0.495)^	<i>b</i> ₃₄ (Q28=No)	7.084 (3.65) ^	<i>b</i> ₆₄ (Q44 = Yes)	0.171 (1.027)
<i>b</i> ₅ (Q5)	0.431 (0.451)	<i>b</i> ₃₅ (Q28=Yes)	3.542 (3.057)	<i>b</i> ₆₅ (Q45 = Yes)	0.417 (1.190)
<i>b</i> ₆ (Q6)	0.386 (0.497)	<i>b</i> ₃₆ (Q29=No)	-1.364 (3.703)	<i>b</i> ₆₆ (Q46 = No)	-3.889 (2.440)
<i>b</i> ₇ (Q7)	-0.788 (0.495)	<i>b</i> ₃₇ (Q29=Yes)	-1.036 (3.37)	<i>b</i> ₆₇ (Q46 = Yes)	-3.739 (2.555)
<i>b</i> ₈ (Q8)	0.213 (0.416)	<i>b</i> ₃₈ (Q30)	-0.554 (0.421)	<i>b</i> ₆₈ (Q47 = No)	0.700 (3.513)
<i>b</i> ₉ (Q9 = Low)	2.761 (4.543)	<i>b</i> ₃₉ (Q31 =2-3h)	0.234 (1.032)	<i>b</i> ₆₉ (Q47 = Yes)	0.881 (3.645)
<i>b</i> ₁₀ (Q9 = Moderate)	0.816 (0.992)	<i>b</i>₄₀(Q31 = 3-4h)	-8.489 (3.531)*	<i>b</i> ₇₀ (Q48 = First thought)	2.663 (2.088)
<i>b</i>₁₁(Q10)	1.202 (0.527)*	<i>b</i> ₄₁ (Q31 = <1h)	1.320 (1.065)	<i>b</i>₇₁(Q48 = Think through)	4.905 (2.039) *
<i>b</i>₁₂(Q11)	-1.76 (0.635)**	<i>b</i> ₄₂ (Q31 = >4h)	2.459 (1.841)	<i>b</i>₇₂(Q49)	1.654 (0.737) *
<i>b</i> ₁₃ (Q12)	-0.725 (0.456)	<i>b</i> ₄₃ (Q32 = Dining hall)	0.353 (1.707)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.008 Random effects $c_i \sim N(0,0)$ Occurrence ratio in data: 0.150</p>	
<i>b</i> ₁₄ (Q13)	0.286 (0.535)	<i>b</i> ₄₄ (Q32 = Restaurants)	-2.400 (1.623)		
<i>b</i> ₁₅ (Q14)	0.151 (0.483)	<i>b</i>₄₅(Q32 = Home)	-2.947 (1.334)*		
<i>b</i>₁₆(Q15)	-1.509 (0.607)*	<i>b</i> ₄₆ (Q33=No)	2.245 (1.366)		
<i>b</i> ₁₇ (Q16)	-0.098 (0.453)	<i>b</i>₄₇(Q33=Some)	5.605 (1.725) **		
<i>b</i> ₁₈ (Q17)	-0.937 (0.573)	<i>b</i>₄₈(Q34)	-1.487 (0.605) *		
<i>b</i> ₁₉ (Q18)	0.617 (0.456)	<i>b</i> ₄₉ (Q35)	-0.572 (0.524)		
<i>b</i> ₂₀ (Q19)	-1.089 (0.588)^	<i>b</i>₅₀(Q36)	0.836 (0.363) *		
<i>b</i>₂₁(Q20 = Over budget)	-4.823 (2.034)*	<i>b</i> ₅₁ (Q37 = >3-4)	1.239 (1.163)		
<i>b</i> ₂₂ (Q20 = Under budget)	1.676 (1.221)	<i>b</i> ₅₂ (Q37 = >4)	-2.386 (1.707)		
<i>b</i> ₂₃ (Q21 = Behind sched.)	-0.657 (1.806)	<i>b</i> ₅₃ (Q37 = None)	1.747 (1.159)		
<i>b</i> ₂₄ (Q21 = On sched.)	-1.068 (1.711)	<i>b</i> ₅₄ (Q38)	-0.696 (0.468)		
<i>b</i> ₂₅ (Q22 = More reqs)	-0.914 (1.534)	<i>b</i>₅₅(Q39=Reqs)	-8.005 (2.318)		
<i>b</i>₂₆(Q22=reqs as planned)	-3.529 (1.54)*	<i>b</i>₅₆(Q39=Sched)	-2.639 (1.217)		
<i>b</i>₂₇(Q23)	1.387 (0.584)*	<i>b</i> ₅₇ (Q40=Reqs)	1.341 (1.771)		
<i>b</i>₂₈(Q24)	2.355 (0.845)**	<i>b</i> ₅₈ (Q40=Sched)	-1.156 (1.239)		
<i>b</i> ₂₉ (Q25)	-0.411 (0.564)	<i>b</i> ₅₉ (Q41 = No)	-4.230 (2.618)		

Table 38: Mixed-effects logistic regression model coefficients for FC6: Inadequately communicated.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
<i>a</i>	7.068 (3.557)	<i>b</i>₃₀(Q26 = No)	-3.264 (1.5)*	<i>b</i> ₆₀ (Q41 = Yes)	0.167 (1.13)
<i>b</i>₁(Q1)	-1.190 (0.351)***	<i>b</i> ₃₁ (Q26 = Yes)	-0.835 (1.302)	<i>b</i>₆₁(Q42 = No)	-3.457 (1.516)*
<i>b</i>₂(Q2)	0.591 (0.253)*	<i>b</i> ₃₂ (Q27=No)	-2.068 (1.419)	<i>b</i>₆₂(Q42 = Yes)	-3.874 (1.573)*
<i>b</i> ₃ (Q3)	0.193 (0.282)	<i>b</i> ₃₃ (Q27=Yes)	-1.015 (1.35)	<i>b</i> ₆₃ (Q43 = Yes)	-0.729 (0.658)
<i>b</i> ₄ (Q4)	-0.121 (0.297)	<i>b</i> ₃₄ (Q28=No)	1.759 (1.863)	<i>b</i> ₆₄ (Q44 = Yes)	-0.854 (0.572)
<i>b</i>₅(Q5)	0.695 (0.291)*	<i>b</i> ₃₅ (Q28=Yes)	3.132 (2.042)	<i>b</i> ₆₅ (Q45 = Yes)	1.496 (0.638)
<i>b</i> ₆ (Q6)	0.107 (0.28)	<i>b</i> ₃₆ (Q29=No)	2.638 (3.175)	<i>b</i> ₆₆ (Q46 = No)	-0.991 (1.124)
<i>b</i> ₇ (Q7)	-0.331 (0.299)	<i>b</i> ₃₇ (Q29=Yes)	2.842 (3.227)	<i>b</i> ₆₇ (Q46 = Yes)	-1.119 (1.102)
<i>b</i> ₈ (Q8)	0.212 (0.26)	<i>b</i> ₃₈ (Q30)	-0.273 (0.243)	<i>b</i> ₆₈ (Q47 = No)	-2.318 (1.411)
<i>b</i> ₉ (Q9 = Low)	-1.025 (1.026)	<i>b</i> ₃₉ (Q31 =2-3h)	-0.2 (0.671)	<i>b</i>₆₉(Q47 = Yes)	-3.122 (1.554)*
<i>b</i>₁₀(Q9 = Moderate)	-1.296 (0.596)*	<i>b</i> ₄₀ (Q31 = 3-4h)	0.156 (1.479)	<i>b</i> ₇₀ (Q48 = First thought)	-0.777 (0.923)
<i>b</i> ₁₁ (Q10)	0.184 (0.246)	<i>b</i> ₄₁ (Q31 = <1h)	-0.946 (0.642)	<i>b</i> ₇₁ (Q48 = Think through)	-0.787 (0.843)
<i>b</i> ₁₂ (Q11)	0.076 (0.272)	<i>b</i> ₄₂ (Q31 = >4h)	0.551 (1.013)	<i>b</i> ₇₂ (Q49)	-0.271 (0.285)
<i>b</i> ₁₃ (Q12)	0.031 (0.269)	<i>b</i>₄₃(Q32 = Dining hall)	-3.366 (1.493)*	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.164 Random effects $c_i \sim N(0, 0.954^2)$ Occurrence ratio in data: 0.200</p>	
<i>b</i> ₁₄ (Q13)	-0.039 (0.273)	<i>b</i> ₄₄ (Q32 = Restaurants)	-0.472 (1.036)		
<i>b</i> ₁₅ (Q14)	-0.339 (0.269)	<i>b</i> ₄₅ (Q32 = Home)	-0.832 (0.882)		
<i>b</i> ₁₆ (Q15)	-0.277 (0.3)	<i>b</i> ₄₆ (Q33=No)	-1.22 (0.703)^		
<i>b</i>₁₇(Q16)	-0.521 (0.251)*	<i>b</i> ₄₇ (Q33=Some)	-0.516 (0.728)		
<i>b</i>₁₈(Q17)	0.571 (0.294)*	<i>b</i> ₄₈ (Q34)	-0.352 (0.288)		
<i>b</i> ₁₉ (Q18)	0.057 (0.28)	<i>b</i>₄₉(Q35)	0.573 (0.283)*		
<i>b</i> ₂₀ (Q19)	-0.29 (0.261)	<i>b</i> ₅₀ (Q36)	0.036 (0.307)		
<i>b</i> ₂₁ (Q20 = Over budget)	0.797 (0.721)	<i>b</i> ₅₁ (Q37 = >3-4)	-0.292 (0.67)		
<i>b</i> ₂₂ (Q20 = Under budget)	-1.045 (0.709)	<i>b</i> ₅₂ (Q37 = >4)	-0.086 (0.792)		
<i>b</i> ₂₃ (Q21 = Behind sched.)	-1.718 (0.963)^	<i>b</i> ₅₃ (Q37 = None)	-0.693 (0.709)		
<i>b</i> ₂₄ (Q21 = On sched.)	-0.961 (0.864)	<i>b</i> ₅₄ (Q38)	0.333 (0.3)		
<i>b</i> ₂₅ (Q22 = More reqs)	-0.588 (0.953)	<i>b</i> ₅₅ (Q39=Reqs)	0.554 (0.712)		
<i>b</i> ₂₆ (Q22=reqs as planned)	-0.458 (0.702)	<i>b</i> ₅₆ (Q39=Sched)	-1.167 (0.641)^		
<i>b</i> ₂₇ (Q23)	-0.294 (0.302)	<i>b</i> ₅₇ (Q40=Reqs)	0.29 (1.275)		
<i>b</i> ₂₈ (Q24)	-0.117 (0.296)	<i>b</i> ₅₈ (Q40=Sched)	0.003 (0.618)		
<i>b</i> ₂₉ (Q25)	-0.543 (0.337)	<i>b</i> ₅₉ (Q41 = No)	0.747 (1.149)		

Table 39: Mixed-effects logistic regression model coefficients for FC7: Subjected to inadequate testing.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
<i>a</i>	-8.212 (5.143)	<i>b</i> ₃₀ (Q26 = No)	5.478 (3.288)^	<i>b</i> ₆₀ (Q41 = Yes)	5.894 (3.053)^
<i>b</i>₁(Q1)	1.328 (0.481)**	<i>b</i> ₃₁ (Q26 = Yes)	3.848 (3.004)	<i>b</i> ₆₁ (Q42 = No)	-0.200 (0.671)
<i>b</i> ₂ (Q2)	-0.772 (0.48)	<i>b</i> ₃₂ (Q27=No)	-2.76 (1.95)	<i>b</i> ₆₂ (Q42 = Yes)	1.516 (0.989)
<i>b</i>₃(Q3)	0.967 (0.408)*	<i>b</i> ₃₃ (Q27=Yes)	-1.64 (1.86)	<i>b</i> ₆₃ (Q43 = Yes)	-0.711 (1.122)
<i>b</i>₄(Q4)	-1.764 (0.555)**	<i>b</i> ₃₄ (Q28=No)	-4.274 (2.585)^	<i>b</i> ₆₄ (Q44 = Yes)	-0.95 (0.896)
<i>b</i> ₅ (Q5)	0.693 (0.449)	<i>b</i> ₃₅ (Q28=Yes)	-4.058 (2.706)	<i>b</i> ₆₅ (Q45 = Yes)	0.391 (1.082)
<i>b</i> ₆ (Q6)	0.675 (0.45)	<i>b</i> ₃₆ (Q29=No)	4.225 (2.564)^	<i>b</i> ₆₆ (Q46 = No)	1.581 (2.614)
<i>b</i> ₇ (Q7)	0.41 (0.472)	<i>b</i> ₃₇ (Q29=Yes)	3.28 (2.787)	<i>b</i> ₆₇ (Q46 = Yes)	1.111 (2.588)
<i>b</i> ₈ (Q8)	-0.773 (0.428)^	<i>b</i>₃₈(Q30)	1.463 (0.529)*	<i>b</i> ₆₈ (Q47 = No)	-1.513 (3.137)
<i>b</i> ₉ (Q9 = Low)	1.38 (1.346)	<i>b</i>₃₉(Q31 =2-3h)	-1.94 (1.129)*	<i>b</i> ₆₉ (Q47 = Yes)	-2.219 (3.356)
<i>b</i> ₁₀ (Q9 = Moderate)	-1.415 (0.952)	<i>b</i> ₄₀ (Q31 = 3-4h)	-3.239 (2.196)	<i>b</i> ₇₀ (Q48 = First thought)	-2.836 (1.834)
<i>b</i> ₁₁ (Q10)	0.304 (0.438)	<i>b</i> ₄₁ (Q31 = <1h)	-0.338 (0.946)	<i>b</i> ₇₁ (Q48 = Think through)	-2.404 (1.517)
<i>b</i> ₁₂ (Q11)	0.717 (0.457)	<i>b</i> ₄₂ (Q31 = >4h)	-2.048 (1.622)	<i>b</i> ₇₂ (Q49)	-0.405 (0.488)
<i>b</i> ₁₃ (Q12)	0.5 (0.427)	<i>b</i> ₄₃ (Q32 = Dining hall)	-0.556 (1.686)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001</p> <p>Median scaled residual: -0.022 Random effects $c_i \sim N(0,0)$ Occurrence ratio in data: 0.150</p>	
<i>b</i> ₁₄ (Q13)	-0.893 (0.477)^	<i>b</i> ₄₄ (Q32 = Restaurants)	0.493 (1.694)		
<i>b</i> ₁₅ (Q14)	0.196 (0.403)	<i>b</i> ₄₅ (Q32 = Home)	0.207 (1.123)		
<i>b</i> ₁₆ (Q15)	-0.25 (0.473)	<i>b</i>₄₆(Q33=No)	3.236 (1.349)*		
<i>b</i>₁₇(Q16)	1.046 (0.506)*	<i>b</i>₄₇(Q33=Some)	2.573 (1.131)*		
<i>b</i> ₁₈ (Q17)	-0.161 (0.472)	<i>b</i> ₄₈ (Q34)	0.327 (0.461)		
<i>b</i> ₁₉ (Q18)	-0.344 (0.36)	<i>b</i> ₄₉ (Q35)	0.819 (0.523)		
<i>b</i>₂₀(Q19)	-1.319 (0.514)*	<i>b</i> ₅₀ (Q36)	-1.684 (1.342)		
<i>b</i> ₂₁ (Q20 = Over budget)	-1.687 (1.388)	<i>b</i> ₅₁ (Q37 = >3-4)	0.41 (1.163)		
<i>b</i> ₂₂ (Q20 = Under budget)	-0.073 (1.138)	<i>b</i> ₅₂ (Q37 = >4)	-0.525 (1.491)		
<i>b</i>₂₃(Q21 = Behind sched.)	-3.712 (1.473)*	<i>b</i> ₅₃ (Q37 = None)	-0.048 (1.212)		
<i>b</i> ₂₄ (Q21 = On sched.)	1.691 (1.158)	<i>b</i> ₅₄ (Q38)	-0.966 (0.502)^		
<i>b</i> ₂₅ (Q22 = More reqs)	-4.117 (3.182)	<i>b</i> ₅₅ (Q39=Reqs)	-1.351 (1.261)		
<i>b</i> ₂₆ (Q22=reqs as planned)	1.651 (1.241)	<i>b</i> ₅₆ (Q39=Sched)	-0.047 (1.023)		
<i>b</i> ₂₇ (Q23)	-0.169 (0.475)	<i>b</i>₅₇(Q40=Reqs)	-6.069 (2.879)*		
<i>b</i> ₂₈ (Q24)	-0.116 (0.529)	<i>b</i>₅₈(Q40=Sched)	-2.935 (1.333)*		
<i>b</i>₂₉(Q25)	1.622 (0.518)**	<i>b</i> ₅₉ (Q41 = No)	4.376 (2.64)^		

Table 40: Mixed-effects logistic regression model coefficients for FC8: Managed risk poorly.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
<i>a</i>	9.811 (7.686)	<i>b</i> ₃₀ (Q26 = No)	0.099 (0.247)	<i>b</i>₆₀(Q42 = No)	-10.712 (4.633)*
<i>b</i> ₁ (Q1)	0.132 (0.599)	<i>b</i> ₃₁ (Q26 = Yes)	-4.839 (2.705)^	<i>b</i>₆₁(Q42 = Yes)	-12.751 (5.461)*
<i>b</i> ₂ (Q2)	-1.144 (0.874)	<i>b</i> ₃₂ (Q27=No)	7.657 (4.349)^	<i>b</i> ₆₂ (Q43 = No)	-2.640 (3.730)
<i>b</i>₃(Q3)	2.233 (0.875)*	<i>b</i>₃₃(Q27=Yes)	8.157 (4.071)*	<i>b</i> ₆₃ (Q43 = Yes)	-0.746 (3.555)
<i>b</i> ₄ (Q4)	0.944 (0.722)	<i>b</i> ₃₄ (Q28=No)	1.598 (4.784)	<i>b</i> ₆₄ (Q44 = Yes)	-0.039 (1.222)
<i>b</i> ₅ (Q5)	-0.5 (0.49)	<i>b</i> ₃₅ (Q28=Yes)	-1.349 (4.119)	<i>b</i> ₆₅ (Q45 = Yes)	-3.095 (1.748)^
<i>b</i> ₆ (Q6)	0.711 (0.683)	<i>b</i> ₃₆ (Q29=No)	0.664 (4.575)	<i>b</i> ₆₆ (Q46 = No)	2.012 (1.341)
<i>b</i> ₇ (Q7)	1.332 (0.824)	<i>b</i> ₃₇ (Q29=Yes)	0.007 (4.17)	<i>b</i> ₆₇ (Q47 = Yes)	-2.708 (1.672)
<i>b</i> ₈ (Q8)	0.967 (0.673)	<i>b</i> ₃₈ (Q30)	-1.012 (0.673)	<i>b</i> ₆₈ (Q48 = First thought)	2.364 (1.871)
<i>b</i> ₉ (Q9 = Low)	-1.238 (3.483)	<i>b</i> ₃₉ (Q31 =2-3h)	-2.285 (1.411)	<i>b</i> ₆₉ (Q48 = Think through)	2.295 (1.524)
<i>b</i> ₁₀ (Q9 = Moderate)	0.606 (1.159)	<i>b</i> ₄₀ (Q31 = 3-4h)	0.784 (2.026)	<i>b</i>₇₀(Q49)	2.562 (1.142)*
<i>b</i> ₁₁ (Q10)	0.347 (0.543)	<i>b</i> ₄₁ (Q31 = <1h)	-0.395 (1.669)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001 Median scaled residual: -0.0135 Random effects $c_i \sim N(0, 0.539^2)$ Occurrence ratio in data: 0.140</p>	
<i>b</i> ₁₂ (Q11)	0.593 (0.684)	<i>b</i> ₄₂ (Q31 = >4h)	-6.732 (4.691)		
<i>b</i> ₁₃ (Q12)	0.333 (0.585)	<i>b</i> ₄₃ (Q32 = Dining hall)	0.805 (2.942)		
<i>b</i>₁₄(Q13)	-1.957 (0.792)*	<i>b</i> ₄₄ (Q32 = Restaurants)	-6.901 (4.022)^		
<i>b</i> ₁₅ (Q14)	-0.829 (0.603)	<i>b</i> ₄₅ (Q32 = Home)	-2.733 (2.467)		
<i>b</i> ₁₆ (Q15)	0.015 (0.732)	<i>b</i> ₄₆ (Q33=No)	-2.783 (1.594)^		
<i>b</i> ₁₇ (Q16)	1.083 (0.784)	<i>b</i> ₄₇ (Q33=Some)	0.704 (1.497)		
<i>b</i> ₁₈ (Q17)	-0.598 (0.711)	<i>b</i> ₄₈ (Q34)	-0.963 (0.671)		
<i>b</i> ₁₉ (Q18)	-0.236 (0.696)	<i>b</i>₄₉(Q35)	1.745 (0.778)*		
<i>b</i> ₂₀ (Q19)	-0.565 (0.608)	<i>b</i> ₅₀ (Q36)	-0.475 (0.643)		
<i>b</i> ₂₁ (Q20 = Over budget)	-2.298 (2.106)	<i>b</i> ₅₁ (Q37 = >3-4)	-0.21 (1.214)		
<i>b</i> ₂₂ (Q20 = Under budget)	1.975 (1.407)	<i>b</i>₅₂(Q37 = >4)	-5.177 (2.335)*		
<i>b</i>₂₃(Q21 = Behind sched.)	-5.752 (2.383)*	<i>b</i> ₅₃ (Q37 = None)	-2.303 (1.827)		
<i>b</i>₂₄(Q21 = On sched.)	-4.271 (1.811)*	<i>b</i> ₅₄ (Q38)	-0.732 (0.587)		
<i>b</i> ₂₅ (Q22 = More reqs)	5.153 (3.046)^	<i>b</i>₅₅(Q39=Reqs)	-5.774 (2.391)*		
<i>b</i> ₂₆ (Q22=reqs as planned)	1.232 (1.489)	<i>b</i> ₅₆ (Q39=Sched)	1.461 (1.465)		
<i>b</i> ₂₇ (Q23)	0.271 (0.696)	<i>b</i> ₅₇ (Q40=Reqs)	3.012 (3.952)		
<i>b</i>₂₈(Q24)	2.395 (0.927)**	<i>b</i> ₅₈ (Q40=Sched)	-1.534 (1.735)		
<i>b</i> ₂₉ (Q25)	-0.252 (0.763)	<i>b</i> ₅₉ (Q41 = Yes)	2.74 (1.637)*		

Table 41: Mixed-effects logistic regression model coefficients for FC10: Did not allow system aspect to stabilize.

Coefficient	Estimate (error)	Coefficient	Estimate (error)	Coefficient	Estimate (error)
a	-1.811 (3.312)	b_{30} (Q26 = No)	-0.221 (1.126)	b_{60} (Q41 = Yes)	1.476 (1.064)
b_1 (Q1)	-0.225 (0.244)	b_{31} (Q26 = Yes)	-0.118 (0.994)	b_{61} (Q42 = No)	-1.302 (1.191)
b_2 (Q2)	0.199 (0.297)	b_{32} (Q27=No)	-0.847 (1.291)	b_{62} (Q42 = Yes)	-1.32 (1.19)
b_3 (Q3)	-0.14 (0.255)	b_{33} (Q27=Yes)	-0.176 (1.199)	b_{63} (Q43 = Yes)	-1.186 (1.626)
b_4 (Q4)	0.322 (0.253)	b_{34} (Q28=No)	-0.311 (1.012)	b_{64} (Q44 = Yes)	-0.544 (0.482)
b_5 (Q5)	-0.061 (0.234)	b_{35} (Q28=Yes)	0.061 (0.696)	b_{65} (Q45 = Yes)	-0.351 (0.556)
b_6 (Q6)	-0.236 (0.278)	b_{36} (Q29=No)	0.366 (1.455)	b_{66} (Q46 = Yes)	0.859 (0.56)
b_7 (Q7)	-0.432 (0.261)^	b_{37} (Q29=Yes)	-0.742 (0.636)	b_{67} (Q47 = Yes)	-0.501 (0.611)
b_8 (Q8)	-0.118 (0.284)	b_{38} (Q30)	-0.065 (0.238)	b_{68} (Q48 = First thought)	1.126 (0.922)
b_9 (Q9 = Low)	-0.835 (1.622)	b_{39}(Q31 =2-3h)	-1.886 (0.668)**	b_{69}(Q48 = Think through)	1.646 (0.828)*
b_{10} (Q9 = Moderate)	-0.985 (0.517)^	b_{40} (Q31 = 3-4h)	0.789 (0.854)	b_{70} (Q49)	0.114 (0.261)
b_{11}(Q10)	0.572 (0.244)*	b_{41} (Q31 = <1h)	-0.885 (0.647)	<p>^ p < .01 * p < .05 ** p < .01 *** p < .001 Median scaled residual: -0.2532 Random effects $c_i \sim N(0, 0.416^2)$ Occurrence ratio in data: 0.205</p>	
b_{12} (Q11)	-0.035 (0.258)	b_{42} (Q31 = >4h)	0.837 (0.962)		
b_{13} (Q12)	0.487 (0.26)^	b_{43} (Q32 = Dining hall)	2.355 (1.538)		
b_{14} (Q13)	-0.077 (0.258)	b_{44} (Q32 = Restaurants)	1.289 (1.635)		
b_{15} (Q14)	-0.083 (0.241)	b_{45}(Q32 = Home)	2.859 (1.43)*		
b_{16} (Q15)	0.37 (0.269)	b_{46} (Q33=No)	-0.224 (0.635)		
b_{17} (Q16)	-0.104 (0.243)	b_{47}(Q33=Some)	1.161 (0.579)*		
b_{18} (Q17)	-0.293 (0.307)	b_{48} (Q34)	-0.226 (0.274)		
b_{19} (Q18)	0.149 (0.294)	b_{49} (Q35)	-0.286 (0.289)		
b_{20} (Q19)	-0.184 (0.248)	b_{50} (Q36)	-0.203 (0.353)		
b_{21} (Q20 = Over budget)	0.352 (0.719)	b_{51} (Q37 = >3-4)	0.002 (0.654)		
b_{22} (Q20 = Under budget)	0.013 (0.544)	b_{52} (Q37 = >4)	-0.62 (0.755)		
b_{23} (Q21 = Behind sched.)	-0.435 (0.878)	b_{53} (Q37 = None)	-0.359 (0.697)		
b_{24} (Q21 = On sched.)	-0.889 (0.832)	b_{54} (Q38)	0.085 (0.293)		
b_{25} (Q22 = More reqs)	-0.189 (0.937)	b_{55} (Q39=Reqs)	0.082 (0.757)		
b_{26} (Q22=reqs as planned)	-0.471 (0.623)	b_{56} (Q39=Sched)	1.083 (0.662)		
b_{27} (Q23)	0.129 (0.283)	b_{57} (Q40=Reqs)	0.353 (1.242)		
b_{28} (Q24)	-0.253 (0.3)	b_{58} (Q40=Sched)	0.597 (0.64)		
b_{29} (Q25)	-0.274 (0.288)	b_{59} (Q41 = No)	0.953 (1.001)		

APPENDIX B. CONTINGENCY TABLES OF FAILURE CAUSES

Appendix B includes the contingency tables for each failure cause i that I used in the statistical test in Section 3.1 to compare failure cause occurrences between the “PBL” and “IND” samples.

Table 42: Contingency table for FC1: Failed to consider a design aspect.

Failure cause 1	“PBL” sample	“IND” sample	Total
<i>Occurrence</i>	10	29	39
<i>Not occurrence</i>	18	3	21
Total	28	32	60

Table 43: Contingency table for FC2: Used inadequate justification.

Failure cause 2	“PBL” sample	“IND” sample	Total
<i>Occurrence</i>	9	11	20
<i>Not occurrence</i>	19	21	30
Total	28	32	60

Table 44: Contingency table for FC3: Failed to form a contingency plan.

Failure cause 3	“PBL” sample	“IND” sample	Total
<i>Occurrence</i>	7	8	15
<i>Not occurrence</i>	21	24	45
Total	28	32	60

Table 45: Contingency table for FC4: Lacked experience.

Failure cause 4	“PBL” sample	“IND” sample	Total
<i>Occurrence</i>	8	15	23
<i>Not occurrence</i>	20	17	37
Total	28	32	60

Table 46: Contingency table for FC5: Kept poor records.

Failure cause 5	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	5	4	9
<i>Not occurrence</i>	23	28	51
Total	28	32	60

Table 47: Contingency table for FC6: Inadequately communicated.

Failure cause 6	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	6	11	17
<i>Not occurrence</i>	22	21	43
Total	28	32	60

Table 48: Contingency table for FC7: Subjected to inadequate testing.

Failure cause 7	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	5	15	20
<i>Not occurrence</i>	23	17	40
Total	28	32	60

Table 49: Contingency table for FC8: Managed risk poorly.

Failure cause 8	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	5	12	17
<i>Not occurrence</i>	23	20	43
Total	28	32	60

Table 50: Contingency table for FC9: Violated procedures.

Failure cause 9	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	3	5	8
<i>Not occurrence</i>	25	27	52
Total	28	32	60

Table 51: Contingency table for FC10: Did not allow system aspect to stabilize.

Failure cause 10	<i>“PBL” sample</i>	<i>“IND” sample</i>	Total
<i>Occurrence</i>	6	16	22
<i>Not occurrence</i>	22	16	38
Total	28	32	60

REFERENCES

- Accreditation Board for Engineering and Technology, Inc. (ABET) (2019) *Criteria for Accrediting Engineering Programs, 2019 – 2020* [online], Available at: <https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2019-2020/#GC3> (Accessed: 14 April 2021).
- Abt, E., Rodricks, J.V., Levy, J.I., Zeise, L. and Burke, T.A. (2010) ‘Science and decisions: advancing risk assessment’, *Risk Analysis*, Vol. 30, No. 7, pp. 1028-1036.
- Al-Bahar, J.F. and Crandall, K.C. (1990) ‘Systematic risk management approach for construction projects’, *Journal of Construction Engineering and Management*, Vol. 116, No. 3, pp. 533-546.
- Aloini, D., Dulmin, R. and Mininno, V. (2007) ‘Risk management in ERP project introduction: Review of the literature’, *Information & Management*, Vol. 44, No. 6, pp. 547-567.
- Aloisio, Diane C. (2019) ‘Lessons from Systems Engineering Failures: Determining Why Systems Fail, the State of Systems Engineering Education, and Building an Evidence-Based Network to Help Systems Engineers Identify and Fix Problems on Complex Projects’, *Purdue University Graduate School. Thesis*.
<https://doi.org/10.25394/PGS.7488569.v1>
- Apostolakis, G.E. (2004) ‘How useful is quantitative risk assessment?’, *Risk Analysis: An International Journal*, Vol. 24, No. 3, pp. 515-520.
- Arlot, S. and Celisse, A. (2010) ‘A survey of cross-validation procedures for model selection’, *Statistics surveys*, Vol. 4, pp. 40-79.
- AS/NZS 4360:1999 (1999) *Risk Management* [online], Available at: http://www.epsonet.eu/mediapool/72/723588/data/2017/AS_NZS_4360-1999_Risk_management.pdf (Accessed: 29 December 2020).
- Atman, C.J., Adams, R.S., Cardella, M.E., Turns, J., Mosborg, S., and Saleem, J. (2007) ‘Engineering design processes: A comparison of students and expert practitioners’, *Journal of Engineering Education*, Vol. 96, No. 44, pp. 359-379.
- Bahill, A. and Dean, F. (1996) ‘What is Systems Engineering? A Consensus of Senior Systems Engineers’, *INCOSE International Symposium*, Vol. 6, No. 1, pp. 500-505.
- Barnard, G.A. (1945) ‘A new test for 2×2 tables’, *Nature*, Vol. 156, p.177.
- Bates, D., Sarkar, D., Bates, M.D. and Matrix, L. (2007) ‘The lme4 package’ [online]. Available at: <ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/packages/lme4.pdf> (Accessed: 24 February 2021).
- Baybutt, P. (2018) ‘The validity of engineering judgment and expert opinion in hazard and risk analysis: The influence of cognitive biases’, *Process Safety Progress*, Vol. 37, No. 2, pp. 205-210.
- Bouti, A. and Kadi, D.A. (1994) ‘A state-of-the-art review of FMEA/FMECA’, *International Journal of reliability, quality and safety engineering*, Vol. 1, No. 4, pp. 515-543.
- Bradsher, K. (2000) *Study of Ford Explorer's Design Reveals a Series of Compromises* [online]. Available at: <https://www.nytimes.com/2000/12/07/business/risky-decision-special-report-study-ford-explorer-s-design-reveals-series.html> (Accessed: 29 December 2020).
- Carbone, T.A. and Tippet, D.D. (2004) ‘Project risk management using the project risk FMEA’, *Engineering Management Journal*, Vol. 16, No. 4, pp. 28-35.

- Carr, V. and Tah, J.H.M. (2001) 'A fuzzy approach to construction project risk assessment and analysis: construction project risk management system', *Advances in engineering software*, Vol. 32, No. 10-11, pp. 847-857.
- Cervone, H.F. (2006) 'Project risk management', *OCLC Systems & Services: International Digital Library Perspectives*, Vol. 22, No. 4, pp. 256-262.
- Charette, R.N. (2008) 'What's wrong with weapons acquisitions?', *IEEE Spectrum*, Vol. 45, No. 11, pp. 33-39.
- Chua, D.K.H., Kog, Y.C. and Loh, P.K. (1999) 'Critical success factors for different project objectives', *Journal of construction engineering and management*, Vol. 125, No. 3, pp. 142-150.
- Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. and Liu, J., (2013) 'A nondegenerate penalized likelihood estimator for variance parameters in multilevel models', *Psychometrika*, Vol. 78, No. 4, pp. 685-709.
- Committee of Sponsoring Organizations of the Treadway Commission (COSO) (2004) *Enterprise risk management-integrated framework: executive summary & framework* [online], Available at: <https://www.coso.org/Documents/COSO-ERM-Executive-Summary.pdf> (Accessed: 29 December 2020).
- Cooper, D., Grey, S., Raymond G., and Walker P. (2005) 'Project risk management guidelines: managing risk in large projects and complex procurements', John Wiley & Sons, Inc., pp. 69.
- Cordery, J.L., Morrison, D., Wright, B.M. and Wall, T.D. (2010) 'The impact of autonomy and task uncertainty on team performance: A longitudinal field study', *Journal of Organizational Behavior*, Vol. 31, No. 2-3, pp. 240-258.
- Cottrell, S. (2006) 'A matter of explanation: assessment, scholarship of teaching and their disconnect with theoretical development', *Medical Teacher*, Vol. 28, No. 4, pp. 305-308.
- DeFillippi, R.J. (2001) 'Introduction: Project-based learning, reflective practices and learning' *Management Learning*, Vol. 32, No. 1, pp. 5-10.
- Denning, S. (2013) *What Went Wrong at Boeing?* [online], Available at: <https://www.forbes.com/sites/stevedenning/2013/01/21/what-went-wrong-at-boeing> (Accessed: 29 December 2020).
- Dietz, T. (2017) 'Drivers of human stress on the environment in the twenty-first century', *Annual Review of Environment and Resources*, Vol. 42, pp.189-213.
- Eccles, D.W. and Tenenbaum, G. (2004) 'Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport', *Journal of Sport and Exercise Psychology*, Vol. 26, No. 4, pp. 542-560.
- Fan, C.F. and Yu, Y.C. (2004) 'BBN-based software project risk management', *Journal of Systems and Software*, Vol. 73, No. 2, pp. 193-203.
- Fang, C. and Marle, F. (2012) 'A simulation-based risk network model for decision support in project risk management', *Decision Support Systems*, Vol. 52, No. 3, pp. 635-644.
- Frank, M., Lavy, I. and Elata, D., (2003) 'Implementing the project-based learning approach in an academic engineering course', *International Journal of Technology and Design Education*, Vol. 13, No. 3, pp.273-288.
- Frosdick, S. (1997) 'The techniques of risk analysis are insufficient in themselves', *Disaster Prevention and Management: An International Journal*, Vol.6, No. 3, pp. 165-177.

- Gandhi, J. and Gorod, A. (2012) 'The Importance of Understanding Systemic Risk in Engineering Management Education', *ASEE Annual Conference and Exposition*, June 10-13, 2012, San Antonio, Texas.
- Georgalis, G. and Marais, K. (2019a) 'Assessment of Project-Based Learning Courses using Crowd Signals', *In ASEE 2019 Annual Conference & Exposition*, Tampa, FL.
- Georgalis, G. and Marais, K. (2019b) 'Can We Use Wisdom-of-the-Crowd to Assess Risk of Systems Engineering Failures?', *In INCOSE International Symposium*, Vol. 29, No. 1, pp. 620-635.
- Georgalis, G. and Marais, K. (2021) 'Predicting failure events from crowd-derived inputs: schedule slips and missed requirements', *In INCOSE International Symposium*, Vol. 31, No. 1.
- Gilson, L.L., Mathieu, J.E., Shalley, C.E. and Ruddy, T.M. (2005) 'Creativity and standardization: complementary or conflicting drivers of team effectiveness?', *Academy of Management journal*, Vol. 48, No. 3, pp. 521-531.
- Halfhill, T., Nielsen, T.M., Sundstrom, E., and Weilbaecher, A. (2005), 'Group personality composition and performance in military service teams', *Military Psychology*, Vol. 17, No. 1, pp.41-54.
- Halinski, R.S. and Feldt, L.S. (1970) 'The selection of variables in multiple regression analysis', *Journal of Educational Measurement*, Vol. 7, No. 3, pp.151-157.
- Hall, W., Palmer, S., and Bennett, M. (2012) 'A longitudinal evaluation of a project-based learning initiative in an engineering undergraduate programme.' *European Journal of Engineering Education*, Vol. 37, No. 2, pp. 155-165.
- Hamilton, B.H., Nickerson, J.A. and Owan, H. (2003) 'Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation', *Journal of political Economy*, Vol. 111, No. 3, pp. 465-497.
- Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E., Robinson, B.S., Hodgson, D.J. & Inger, R. (2018) 'A brief introduction to mixed effects modelling and multi-model inference in ecology', *PeerJ*, Vol. 6, p. e4794.
- Hillson, D. (2014) 'How risky is your project — And what are you doing about it?', *PMI® Global Congress 2014—North America*, Phoenix, AZ.
- Hmelo-Silver, C.E. (2004) 'Problem-based learning: What and how do students learn?', *Educational Psychology Review*, Vol. 16, No. 3, pp. 235-266.
- Hosmer, D.W., Jovanovic, B. and Lemeshow, S. (1989) 'Best subsets logistic regression', *Biometrics*, pp.1265-1270.
- Ika, L.A. (2009) 'Project success as a topic in project management journals', *Project Management Journal*, Vol. 40, No. 4, pp. 6-19.
- Institution of Civil Engineers (2014) *Risk Analysis and Management for Projects (RAMP)*. 3rd ed. ICE Publishing.
- ISO 31000:2018 (2018) *Risk Management—Principles and Guidelines* [online], Available at: <https://www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en> (Accessed: 29 December 2020).
- John, W.T., and Thomas, W. (2000) *A review of research on project-based learning*, The Autodesk Foundation, San Rafael, California.
- Judge, T.A. and Bono, J.E. (2000) 'Five-factor model of personality and transformational leadership', *Journal of applied psychology*, Vol. 85, No. 5, pp. 751.
- Kaufman, G.G. and Scott, K.E. (2003) 'What is systemic risk, and do bank regulators retard or contribute to it?', *The Independent Review*, Vol. 7, No. 3, pp. 371-391.

- Keizera, J.A., Halman, J.I. and Song, M. (2002) 'From experience: applying the risk diagnosing methodology', *Journal of product innovation management*, Vol. 19, No. 3, pp. 213-232.
- Kirkman, B.L. and Rosen, B. (1999) 'Beyond self-management: Antecedents and consequences of team empowerment', *Academy of Management journal*, Vol. 42, No. 1, pp. 58-74.
- Kokotsaki, D., Menzies, V. and Wiggins, A., (2016) 'Project-based learning: A review of the literature', *Improving schools*, Vol. 19, No. 3, pp.267-277.
- Kremljak, Z. and Kafol, C. (2014) 'Types of risk in a system engineering environment and software tools for risk analysis', *Procedia Engineering*, Vol. 69, pp. 177-183.
- Lawless, J.F. and Singhal, K. (1987) 'ISMOD: an all-subsets regression program for generalized linear models I. Statistical and computational background' *Computer Methods and Programs in Biomedicine*, Vol. 24, No. 2, pp.117-124.
- Lee, E., Park, Y. and Shin, J.G. (2009) 'Large engineering project risk management using a Bayesian belief network', *Expert Systems with Applications*, Vol. 36, No. 3, pp. 5880-5887.
- Lehmann, M., Christensen, P., Du, X. and Thrane, M., (2008) 'Problem-oriented and project-based learning (POPBL) as an innovative learning strategy for sustainable development in engineering education', *European journal of engineering education*, Vol. 33, No. 3, pp.283-295.
- Lehner, P., Seyed-Solorforough, M.M., O'Connor, M.F., Sak, S. and Mullin, T. (1997) 'Cognitive biases and time stress in team decision making', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 27, No. 5, pp. 698-703.
- Lim, C.S. and Mohamed, M.Z. (1999) 'Criteria of project success: an exploratory re-examination', *International journal of project management*, Vol. 17, No. 4, pp. 243-248.
- Lineberger, R. and Hussein, A. (2016) *Program Management in Aerospace and Defense Still Late and Over Budget* [online]. Available at: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/us-manufacturing-program-management-aerospace-defense.pdf> (Accessed: 29 December 2020).
- Mato A.S. and Andrés A.M. (1997) 'Simplifying the calculation of the P-value for Barnard's test and its derivatives', *Statistics and Computing*, Vol. 7, No. 2, pp. 137-143.
- McCormick, A.C., Kinzie, J. and Gonyea, R.M. (2013) 'Student engagement: Bridging research and practice to improve the quality of undergraduate education', In *Higher education: Handbook of theory and research* (pp. 47-92). Springer, Dordrecht.
- Marais, K., Saleh, J.H. and Leveson, N.G. (2006) 'Archetypes for organizational safety', *Safety Science*, Vol. 44, No. 7, pp. 565-582.
- Mills, J.E. and Treagust, D.F. (2003) 'Engineering education—Is problem-based or project-based learning the answer', *Australasian Journal of Engineering Education*, Vol. 3, No. 2, pp. 2-16.
- Montibeller, G. and Von Winterfeldt, D. (2015) 'Cognitive and motivational biases in decision and risk analysis. Risk Analysis, Vol. 35, No. 7, pp. 1230-1251.
- Muriana, C. and Vizzini, G. (2017) 'Project risk management: A deterministic quantitative technique for assessment and mitigation', *International Journal of Project Management*, Vol. 35, No. 3, pp. 320-340.
- Mundlak, Y. (1978) 'On the pooling of time series and cross section data', *Econometrica: Journal of the Econometric Society*, Vol. 46, No. 1, pp. 69-85.

- Nan, N. and Harter, D.E. (2009) 'Impact of budget and schedule pressure on software development cycle time and effort' *IEEE Transactions on Software Engineering*, Vol. 35, No. 5, pp.624-637.
- Nolan, A., Pickard, A.C., Nolan, J., Beasley, R. and Pruitt, T.C. (2018) 'How Many Systems Engineers Does It Take to Change a Light Bulb?', *In INCOSE International Symposium*, Vol. 28, No. 1, pp. 777-790.
- Paté-Cornell, M.E. (1993) 'Learning from the piper alpha accident: A postmortem analysis of technical and organizational factors' *Risk Analysis*, Vol. 13, No. 2, pp. 215-232.
- Peeters, M.A., Van Tuijl, H.F., Rutte, C.G. and Reymen, I.M. (2006) 'Personality and team performance: a meta-analysis', *European Journal of Personality: Published for the European Association of Personality Psychology*, Vol. 20, No. 5, pp. 377-396.
- Pinto, J.K. and Slevin, D.P. (1987) 'Critical factors in successful project implementation', *IEEE transactions on engineering management*, Vol. 1, pp. 22-27.
- Project Management Institute (2020) *Pulse of the Profession 2020: Ahead of the Curve* [online]. Available at: <https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pmi-pulse-2020-appendix.pdf?v=f3ef13d4-2187-4818-a80e-dd8045157d97> (Accessed: 29 December 2020).
- Rasmussen, J. (1997) 'Risk management in a dynamic society: a modelling problem', *Safety science*, Vol. 27, No. 2-3, pp. 183-213.
- Raz, T., Shenhar, A.J. and Dvir, D. (2002) 'Risk management, project success, and technological uncertainty', *R&D Management*, Vol. 32, No. 2, pp. 101-109.
- Reagans, R., Argote, L. and Brooks, D. (2005) 'Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together', *Management science*, Vol. 51, No. 6, pp. 869-881.
- Röhm J. and Mansmann U. (1999) 'Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority', *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Vol. 41, No. 2, pp. 149-170.
- Rockenbach, B., Sadrieh, A. and Mathauschek, B. (2007) 'Teams take the better risks', *Journal of Economic Behavior & Organization*, Vol. 63, No. 3, pp. 412-422.
- Salas, E., Cooke, N.J. and Rosen, M.A. (2008) 'On teams, teamwork, and team performance: Discoveries and developments', *Human factors*, Vol. 50, No. 3, pp. 540-547.
- Savery, J.R. (2006) 'Overview of problem-based learning: Definitions and distinctions' *The Interdisciplinary Journal of Problem-based Learning*, Vol. 1, No. 1, pp. 9-20.
- Schmidheiny, K. (2016) 'Panel data: fixed and random effects', *Short Guides to Microeconometrics*, University of Basel[online]. Available at: <https://www.schmidheiny.name/teaching/panel2up.pdf> (Accessed: 24 February 2021).
- Shenhar, A.J., Tishler, A., Dvir, D., Lipovetsky, S. and Lechler, T. (2002) 'Refining the search for project success factors: a multivariate, typological approach', *R&D Management*, Vol. 32, No. 2, pp. 111-126.
- Sjöberg, L. (2000) 'Factors in risk perception', *Risk analysis*, Vol. 20, No. 1, pp. 1-12.
- Sorenson, D. and Marais, K. (2016) 'Patterns of causation in accidents and other systems engineering failures', *In Systems Conference (SysCon), 2016*, Annual IEEE (pp. 1-8).
- Suissa S. & Shuster J. (1985) 'Exact unconditional sample sizes for the 2× 2 binomial trial', *Journal of the Royal Statistical Society. Series A (General)*, pp. 317-327.

- Tah, J.H.M. and Carr, V. (2001) 'Towards a framework for project risk knowledge management in the construction supply chain', *Advances in Engineering Software*, Vol. 32, No. 10-11, pp. 835-846.
- Takahashi, D. (2008) *Xbox 360 defects: an inside history if Microsoft's video game console woes* [online]. Available at: <https://venturebeat.com/2008/09/05/xbox-360-defects-an-inside-history-of-microsofts-video-game-console-woes/> (Accessed: 29 December 2020).
- U.S. Government Accountability Office (GAO) (2017), *F-35 Joint Strike Fighter; DOD Needs to Complete Developmental Testing Before Making Significant New Investments, Report to Congressional Committees* [online]. Available at: <https://www.gao.gov/assets/690/684207.pdf> (Accessed: 29 December 2002).
- van Mierlo, H., Rutte, C.V., Vermunt, J.K., Kompier, M.A.J. and Doorewaard, J.A.M.C. (2006), 'Individual autonomy in work teams: The role of team autonomy, self-efficacy, and social support', *European Journal of Work and Organizational Psychology*, Vol. 15, No. 3, pp. 281-299.
- Virgă, D., Curşeu, P.L., Maricuţoiu, L., Sava, F.A., Macsinga, I. and Măgurean, S. 2014, 'Personality, relationship conflict, and teamwork-related mental models', *PloS one*, Vol. 9, No. 11, pp.e110223.
- Wallace, L., Keil, M. and Rai, A. (2004) 'Understanding software project risk: a cluster analysis', *Information & Management*, Vol. 42, No. 1, pp. 115-125.
- White, D. and Fortune, J. (2002) 'Current practice in project management—An empirical study', *International journal of project management*, Vol. 20, No. 1, pp. 1-11.
- Williams, T.M. (1996) 'The two-dimensionality of project risk', *International Journal of Project Management*, Vol. 14, No. 3, pp. 185-186.
- Yoon, Y., Tamer, Z. and Hastak, M. (2014) 'Protocol to enhance profitability by managing risks in construction projects', *Journal of Management in Engineering*, Vol. 31, No. 5, pp. 04014090.
- Zwikael, O. and Ahn, M. (2011) 'The effectiveness of risk management: an analysis of project risk planning across industries and countries', *Risk analysis*, Vol. 31, No. 1, pp. 25-37.