E1a: Understanding Threats

Experiment Reproducibility Report

Security Experiments and Measurements

Vrije Universiteit 2022

Subject ID: _______2722930________
Subject Name and Surname: ______________Haohui Zhang_______________
VU email: __h17.zhang@subject.vu.nl__

Plagiarism statement

I hereby confirm that this assignment is my own work, is not copied from any other person's work (published or unpublished) and was not shared with any other person that could copy its contents.

Signature:

___________________**Haohui Zhang**__________________

____

Experiment Description

1. Intervention

At the time of the experiment you didn't know what was the difference between the intervention group and the control group. By the time you write this report you will know. You should describe here briefly what the intervention is and why this is interesting.

In this report, we describe a controlled experiment with 86 master students (43 in group A and 43 in group B), who were asked to distinguish the real threats from 11 potential threats. In group A, subjects received a branch of material of the app deployment in Github including a DFD, while on the contrary, subjects in group B received the material without DFD. Ergo, the intervention in this experiment is whether the group is provided with the DFD, and group B is the control group. The purpose of this experiment is to find out whether the DFD is in fact useful and significantly helps in understanding the security threats of a system. DFD is the acronym of the Data Flow Diagram which is a way of representing a flow of data through a process or a system. What interesting is that the DFD is proven very practical to understand and distinguish the security threats of the system. Thus, this intervention will cause considerable accuracy difference of two groups.

2. Metric of Success

Explain how you would measure the success of the intervention using the data that would be given you. This is not the analysis which will happen later. Here you just explain why this is a good choice in your opinion.

The raw data is composed by three parts, respectively training raw data, understanding of threats raw data and the cleaned understanding of threats raw data of the last year experiments. The training raw data contains the group separation information and whether the participants agree their anonymized answers can be used for educational and research purposes. The group separation information is the most important data in this experiment before it represents the intervention directly. Also, out of respect for the participants, the future analysis will not conclude the participants data who didn't successfully take the training. The last year raw data is the background investigation of the subjects which has nothing to do with the success of the intervention of this experiment. It only provides the right answer of the 11 experiment questions. After we completed the initial data filtering, we have identified 41 valid data points for group A and 39 valid data points for group B.

The training raw data composed by the time statistics on the solving the problem and the participants' self-knowledge level assessment. The understanding of threats raw data can be separated into three parts, the time statistics on the solving the problem, the participants answers, and the participants' self-evaluation of the experiment and the confidence of their results. Note that no matter whether the subjects know more about the software design techniques or models etc., it has no significant effects on the intervention (DFD), therefore, the background knowledge on software design model cannot influentially affect the controlled experiment analysis. On the contrary, based on the practical experience, the knowledge on STRIDE, Github and DFD may efficiently help the subject to find out the real threats. Therefore, we also need to find the correlation between the subject's knowledge level and the accuracy of experiment results.

The purpose of this section is to determine the intervention succeed or not. To measure the success of the intervention, firstly, it needs to delete the data without consent to use. Then, it needs to separate the understanding of threats raw data based on the group information. After that, it needs to calculate the accuracy of each participant answers and the precision of the overall experiment results. Besides, the true positive rate, true negative rate, negative predictive rate, false positive rate, false negative rate and false discovery rate are also needed to calculate. Since the metric of success should be determined in a global way, the accuracy only shows the intervention impacts on each participant locally, we will use the global performance metrics, precision and recall, as the measure of success. The mean and standard deviation of subject accuracy and overall accuracy will also be referenced.

3. Subjects & Process

Describe the background of the experiment subjects (who participated in the experiment). Describe the process that was used to perform the experiment in a replicable manner (pay attention to details, e.g. what were the exact steps, is randomization of steps present, etc.)

The background of the experiment subjects are all computer science subjects. Roughly all of the subjects had very limited knowledge of all the context of the secure design topics prior to the experiment, this kind of subjects were randomly and uniformly distributed in two groups. Only very few subjects have a very good prior knowledge who are distributed in group B. Although the different background of the subjects will definitely influence the experiment results, according to the statistics, the bias can be ignored.

	Group A	Group B
Experiment data	43	43
Valid data	41	39

TABLE 1

The main step of the experimental process is the execution of the comparison. The replicable process is specifically as follow:

1. The first step is grouping experiment subjects. The teacher separated all the subjects into two group randomly.

2. The second step is conducting intervention. In this step, subjects in different groups received different material which is specifically whether or not including the DFD. Besides, the teacher controlled both group having same reading time and tried to make sure all the subjects have sufficient time to read the material and have a clear understanding of what the task asked them to do.

3. The third step is carrying out the experiments. After reading and watching all the material, the relevant knowledge background of the subjects was firstly investigated. Then, the subjects were asked to distinguish 11 potential threats whether they are real threats within a limited time. After the submitting the results, all subjects were requested to have a self-evaluation on their results. This evaluation is also very important to the experiment because it reflects the subjects' knowledge of the material and the subject's intuitive perception of the experimental questions and their answers.

4. The forth step is collecting the data. All the answers are recorded, the related time is recorded, and all the relevant information is aggregated. All data are aggregated into three excel files according to the subject's subject number and category.

5. The fifth step is filtering and visualizing the data. We performed data cleaning on the data in three excel, specifically deleting the dirty data based on the experiment time, and whether they had a clear understanding of what the task asked you to do and whether they took the training procedure. After that, we purposed various statistics measures for evaluation, plotted the data for easier comparison and elaborated the results quantitatively and qualitatively.
6. The last step is analysing the results and generating the conclusion. We conducted the analysis procedure by using a statistical test. After that, we analyzed the subjective data of all subjects and compared it with the results of our test. At last, we generated the conclusion.

4. Discussion and Limitations

Reflect on this experiment and describe what are the key criticalities and limitations of this particular experiment design. The discussed limitations and critical points must be specific to the experiment (e.g. identified co-founding variables), and not general limitations of experimentation.

The key criticalities of this experiment design are the intervention implementation and the determination of the success metric. These two factors are so crucial because if the intervention and control of the experiment fail, the validity and usefulness of collected data cannot be compromised. All the effective analysis is based on a correct implementation of an intervention. Otherwise, all the data is dirty data and is considered to be of poor quality. For example, if some subjects in group B (control group) also receive the experimental treatment, getting the DFD of the Github during the experiment, those subjects' answer will be invalid and must be eliminated from the analysis. Such occurrences can adversely impact the overall distribution and randomness of the data and ultimately lead to higher error rates.

As for the determination of the success metric, it is essential to the validity of experimental results, because the experimental conclusions we get are based on valuable metrics. If metrics have no value, the experimental conclusions are meaningless. The careful consideration of success metrics is especially important in this study, which has opted to use precision and recall instead of accuracy as a success metric. This is because accuracy can only reflect everyone's situation, while the precision and recall can reflect the subject's performance. Besides, precision and recall are more effective in capturing subject-specific performance as compared to accuracy, which may only provide a general overview of performance. In the same way, the mean of accuracy has more utility than the basic accuracy.

The key limitations of this experiment design are the subjects' background and experiment time. As an uncontrolled aspect, the subjects' background or experience may have an impact on the experimental results. If the subject is very familiar with the architecture of the Github as well as the secure design techniques, this questions in the experiment will be very easy for them to answer, and they will get correct answers very easily. This part of data will definitely bias the experiment results especially when those subjects are not evenly distributed between the two groups. On the contrary, if the subject has no prior knowledge about the Github and the secure design techniques, the time spending on the solving the problem will increase. Although this experiment used a counterbalanced design to mitigate this threat, the subjects were not completely equally. For instances, as in TABLE 2, group A as a whole has more knowledge on Github than group B, while some subjects in group B master the secure design techniques but all subjects in group A are novices. The worst situation is the subject doesn't know what the task asked to do from the start to finish, if so, there is a highly probability all the data collected about him was filled out blindly.

Group	Experience with secure design techniques (e.g., STRIDE, DFD)			
	Novice	Taken lectures	Tried tools	Profession
Group A	35	4	0	0
Group B	38	1	1	1

Group	Experience with Github			
	Novice	Taken lectures	Tried tools	Profession
Group A	14	9	11	5
Group B	15	13	13	0

TABLE 2

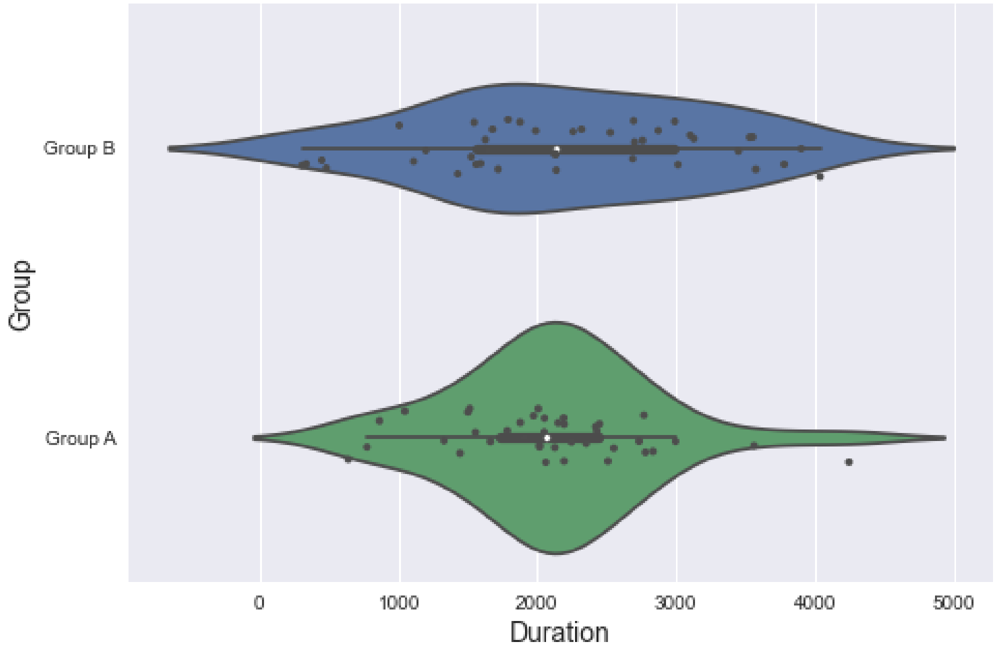


FIGURE 1

Duration time is also very important, if the subject takes too long or the spending time exceeds the limit, there is a high probability that the subject gave the answer to the question by cheating or searching on the Internet. Such data is also invalid because they are not in the prescribed intervention and will also bias the result. Fortunately, in our experiments, all the valid experiment are finished in 5 minutes to 71 minutes, which means all the data are reasonable in the duration time scale. Figure 1 shows violin plots with the time in seconds to implement the proposed tasks of two groups.

Analysis Description

5. Descriptive Statistics

Present key characteristics of the data you are analyzing using some basic descriptive statistics measures (e.g. mean, standard deviation of subjects or other metrics of the intervention). Add graphics is appropriate.

There are 11 potential threats separately:
1. EXPLOIT-REMOTE-REPO
2. LEAKED-CONFIG-FILE
3. STOLEN-AUTH-INFO
4. DOS-SERVER
5. MALICIOUS-CODE-GITHUB
6. ELEVATION-PRIVILEDGED-ACCESS
7. DOS-REMOTE-REPO
8. DISCLOSE-THIRD-PARTY
9. ELEVATION-PRIVILEDGED-REPO
10. ELEVATION-PRIVILEDGED-CODE
11. EXPLOIT-HTTP-PROTOCOL
Six of them are real threats which are the first six.

5.1 Accuracy Per Question

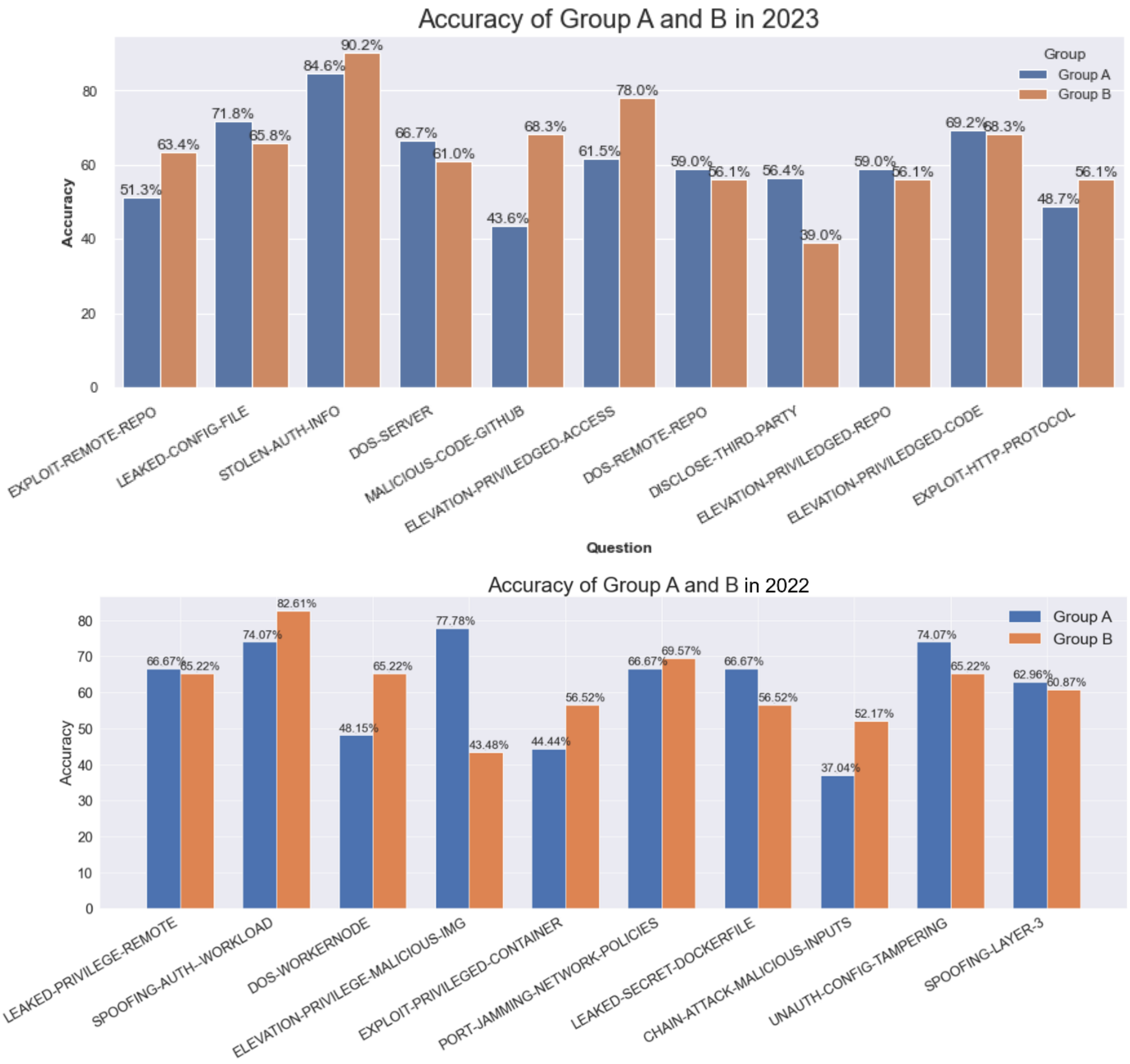
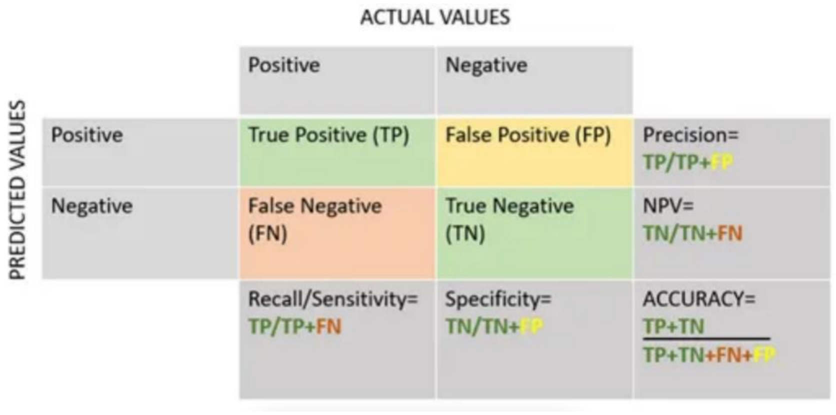


FIGURE 2

This FIGURE 2 shows the accuracy of each question in two years experiment, the scope of the graph is limited to two groups A and B. For example, for the EXPLOIT-REMOTE-REPO threat, there are 20 subjects in group A figuring it out. Because the total amount of subjects in group A is 39, thus, we get this threat's accuracy is 51.3%. For the EXPLOIT-HTTP-PROTOCOL threat, it is not a real threat in our case, in group B, there are 18 subjects mistakenly recognizing it as a real threat, and total amount of subjects in group B is 41, so we calculate (41-18)/41, which is about 56.1%.

An analysis of the data showed that there were differences in the accuracy of the subjects in groups A and B on various questions. From the FIGURE 2, we can clearly observe that the accuracy of each question for group B is better than accuracy for group A as a whole. Especially in the MALICIOUS-CODE-GITHUB and ELEVATION-PRIVILEDGED-ACCESS threats, group B is more than 15 percent more accurate. Specifically, on five of the questions, the subjects in group B demonstrated significantly higher accuracy compared to those in group A. On the other hand, for three of the questions, the subjects in group A performed significantly better than those in group B. For the remaining three questions, the average accuracy of the subjects in the two groups was found to be comparable. These results suggest that there were differences in the ability of the subjects in groups A and B to accurately respond to certain questions, with group B showing better performance on most questions, while group A demonstrated better performance on others. Comparing with the data obtained from last year, the oscillation index is clearly decreased, which can tell us, on the index of accuracy of each question, subjects in group B performed better. These findings highlight the importance of carefully considering the performance of subjects on specific questions, rather than solely relying on overall accuracy scores, in interpreting the results of the experiment.

5.2 Statistic Measures Descriptions



ACTUAL VALUES

		Positive	Negative	
PREDICTED VALUES	Positive	True Positive (TP)	False Positive (FP)	Precision= TP/TP+FP
	Negative	False Negative (FN)	True Negative (TN)	NPV= TN/TN+FN
		Recall/Sensitivity= TP/TP+FN	Specificity= TN/TN+FP	ACCURACY= $\frac{TP+TN}{TP+TN+FN+FP}$

FIGURE 3

True positive, false positive, true negative, and false negative as shown in FIGURE 3 are the four possible outcomes of a binary classification problem where the goal is to classify items into two categories (positive and negative). In the context of a computer security controlled experiment, these outcomes would represent whether a subject has correctly identified a security threat or not.

1. True Positive (TP): When the subject identifies a security threat, and the threat is indeed present, it is considered a true positive.
2. False Positive (FP): When the subject identifies a security threat, but no threat is present, it is considered a false positive.
3. True Negative (TN): When the subject identifies a non-security threat, and no threat is present, it is considered a true negative.

4. False Negative (FN): When the subject fails to identify a security threat, and the threat is present, it is considered a false negative.

Except for the accuracy, we can use the above four outcomes to calculate various performance metrics for a computer security controlled experiment. The following statistics measures can be used to evaluate the effectiveness of a security system:

1. Precision: The precision (positive predictive value) measures the proportion of true positive events (correctly identified security threats) among all events identified as positive by the subject. It represents the accuracy of the subject in identifying real security threats. Specifically, the precision rate reflects how many hits are made in the limited selection. When the precision is 1, it means that all the threats found by the experiment subject are real threats. This indicator can quantitatively reflect the accuracy of a subject to find a threat, in other words, the success rate of a person finding all the threats.
2. Recall: The recall (true positive rate) measures the proportion of true positive events identified by the subject among all actual security threats. It represents the ability of the subject to detect all security threats. A recall of 1 indicates that the subject has found all real threats. The higher the recall value, the better the subject is at correctly identifying all security threats.
3. Specificity: The specificity (true negative rate) measures the ability of the subject to correctly identify non-security threats by measuring the proportion of true negative events (correctly identified as non-security threats) among all events identified as negative by the subject. The higher the specificity value, the better the subject is at correctly identifying non-security threats.
4. Negative predictive value: The negative predictive value measures the proportion of true negative events among all events identified as negative by the subject. It represents the ability of the subject to correctly identify non-security threats and avoid false alarms. The higher the negative predictive value, the better the subject is at correctly identifying non-security threats while minimizing false alarms.
5. False positive rate: The false positive rate (fall-out) measures the proportion of true negative events that are incorrectly identified as positive by the subject. It represents the number of false alarms raised by the subject. The higher the fall-out value, the worse the subject is at minimizing false alarms.
6. False negative rate: The false negative rate measures the proportion of true positive events that are incorrectly identified as negative by the subject. It represents the number of missed security threats. The higher the false negative rate value, the worse the subject is at correctly identifying security threats.
7. False discovery rate: The false discovery rate measures the proportion of false positive events among all events identified as positive by the subject. It represents the number of events flagged as security threats that are not real threats. The higher the false discovery rate value, the worse the subject is at correctly identifying security threats.
8. Overall accuracy: The overall accuracy measures the proportion of correct distinguishes made by the subject among all distinguishes. It represents the overall effectiveness of the subject in detecting security threats and avoiding false alarms. The higher the overall accuracy value, the better the subject is at correctly identifying security threats while minimizing false alarms.
9. Error rate: The error rate measures the proportion of incorrect predictions made by the subject among all distinguishes. It represents the overall effectiveness of the subject in

identifying security threats and avoiding false alarms. The higher the error rate value, the worse the subject is at correctly identifying security threats while minimizing false alarms.

10. F1 Score: The F1 Score is the harmonic mean of precision and recall, which provides a balance between the two measures. It represents the effectiveness of the subject in identifying real security threats while minimizing the number of false alarms. The F1 Score is useful in cases where precision and recall have similar importance in evaluating the subject's performance. It provides a single value that balances both measures, representing the subject's overall ability to detect security threats while avoiding false alarms. A higher F1 Score indicates better performance in identifying real security threats while minimizing false alarms.

5.3 Measurements Per Group

Group A					
Metric	Formula	Mean	Std	Min	Max
Precision	=TP/(TP+FP)	0.656782	0.175080	0.333333	1.000000
Recall	=TP/(TP+FN)	0.632479	0.206543	0.166667	1.000000
Specificity	=TN/(TN+FP)	0.584615	0.235683	0.000000	1.000000
Negative predictive value	=TN/(TN+FN)	0.581516	0.183031	0.250000	1.000000
Fall-out	=FP/(FP+TN)	0.415385	0.235683	0.000000	1.000000
False negative rate	=FN/(TP+FN)	0.367521	0.206543	0.000000	0.833333
False discovery rate	=FP/(TP+FP)	0.343218	0.175080	0.000000	0.666667
Accuracy	=(TP+TN)/(TP+FP+FN+TN)	0.610723	0.153187	0.363636	0.909091
Error rate	=(FP+FN)/(TP+FP+FN+TN)	0.389277	0.153187	0.090909	0.636364
F1 Score	=2*Precision*Recall/ (Precision + Recall)	0.630686	0.161605	0.222222	0.923077

TABLE 3

Group B					
Metric	Formula	Mean	Std	Min	Max
Precision	=TP/(TP+FP)	0.657032	0.158663	0.333333	1.000000
Recall	=TP/(TP+FN)	0.711382	0.207580	0.333333	1.000000
Specificity	=TN/(TN+FP)	0.551220	0.231432	0.000000	1.000000
Negative predictive value	=TN/(TN+FN)	0.629167	0.247757	0.200000	1.000000
Fall-out	=FP/(FP+TN)	0.448780	0.231432	0.000000	1.000000
False negative rate	=FN/(TP+FN)	0.288618	0.207580	0.000000	0.666667
False discovery rate	=FP/(TP+FP)	0.342968	0.158663	0.000000	0.666667
Accuracy	=(TP+TN)/(TP+FP+FN+TN)	0.638581	0.184624	0.272727	1.000000
Error rate	=(FP+FN)/(TP+FP+FN+TN)	0.361419	0.184624	0.000000	0.727273
F1 Score	=2*Precision*Recall/ (Precision + Recall)	0.677312	0.171514	0.333333	1.000000

TABLE 4

TABLE 3 shows the mean, standard deviation, minimum and maximum values of the above 10 metrics of group A as a whole. TABLE 4 shows the mean, standard deviation, minimum and maximum values of the above 10 metrics of group B as a whole.

Based on drawing the comparison between TABLE 3 and TABLE 4, we can observe that subjects in group A and group B performed similarly in terms of precision, which measures the proportion of true security threats among all events identified as positive by the subjects. However, when it comes to recall, which measures the proportion of true security threats identified by the subjects among all actual security threats, the subjects in group B performed better than those in group A. Therefore, the average F1 Score, which is the harmonic mean of precision and recall, was higher in group B than in group A

As for the overall accuracy, subjects in group B also perform better. However, the subjects in group A demonstrated a higher specificity, which measures the proportion of true negative events (correctly identified non-security threats) among all events identified as negative by the subjects. This suggests that subjects in group A were better at correctly identifying non-security threats. Overall, although subjects in group B demonstrated higher recall and overall accuracy, the subjects in group A performed better in terms of specificity and fallout.

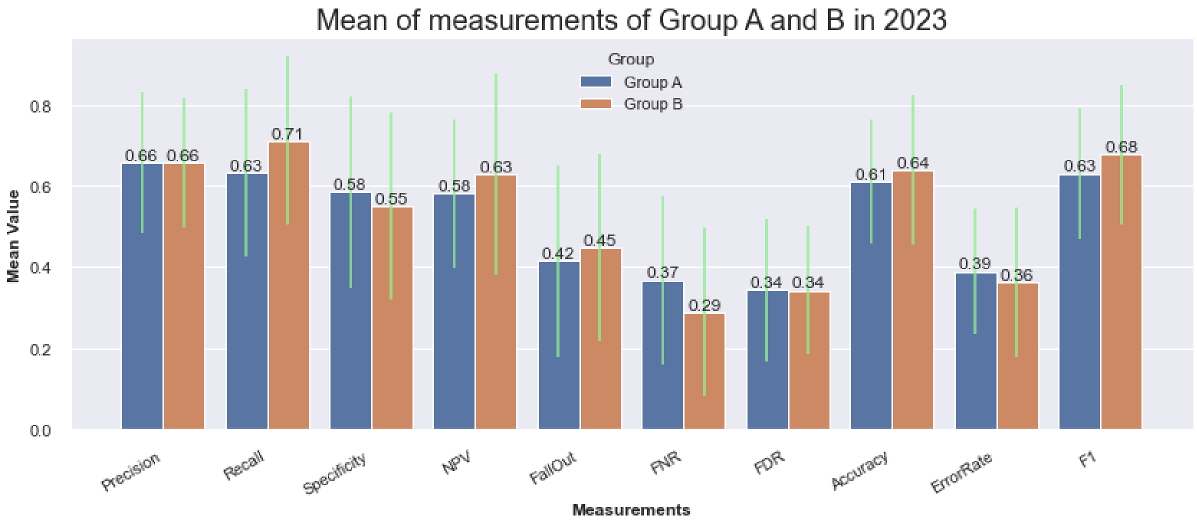


FIGURE 3

FIGURE 3 shows the mean and standard deviation of 10 metrics for two groups. This figure is more illustrative than the above two tables. Intuitively, this figure proves the analysis above. Although the mean precision in group A are similar than in group B, It is unreasonable to only use precision as a evaluation indicator. For instance, one subject only recognized one threat and hit; in this situation his precision was calculated as 1. Although this value is very high, it cannot effectively reflect the ability of the subject of this experiment. This behavior is equivalent to being lazy or cheating. But since we don't know the motive of this experiment subject, maybe he just thinks that threat is the only real threat, so we can't delete his data directly. Therefore, we also need to refer to the recall rate for the overall analysis. Recall is the fraction of relevant instances that were retrieved. In our experiment, the recall rate is a good indicator of whether a person can find all the threats. A recall of 1 indicates that the subject has found all real threats. However, this indicator also has certain problems. If an experimental subject identifies all potential threats as real threats, the recall rate of the experimental subject is 1, but the precision is only 0.5. Therefore, we need to combine the overall observation of precision and recall. Only when both indicators are very high can we truly reflect the ability of the experimental subject to find the real threat. In my view, both precision and recall are very important. Higher precision means that after finding the threat, it can save more time when troubleshooting threats; higher recall means that more threats and vulnerabilities can be found which can directly reduce the possibility of future intrusions. Therefore, F1 score is considered a better metric than precision and recall alone. It takes both metrics into account, providing a balanced evaluation of a subject performance in detecting security threats. Besides, relying on precision or recall alone may result in a biased assessment of a subject's performance. This trade-off between precision and recall can make it challenging to evaluate the effectiveness of a subject in identifying real security threats while minimizing false alarms.

Based on the results of all performance metrics, we can conclude that subjects in group B were better at correctly identifying security threats while minimizing false alarms compared to those in group A, while subjects in group A were better at correctly identifying non-security threats.

5.4 Measurements Per Subject

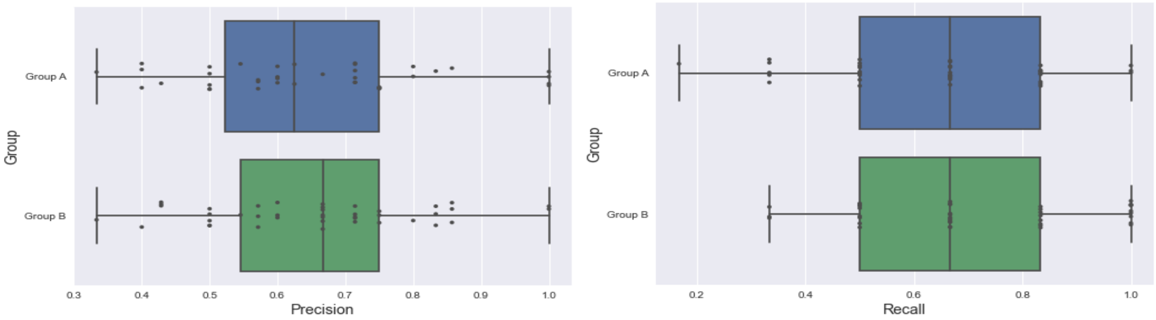


FIGURE 4

This boxplot figure reflects the precision and recall of each subject in group A and group B. This graph is counted and sorted according to the TP, TN, FP and FN of each subject.

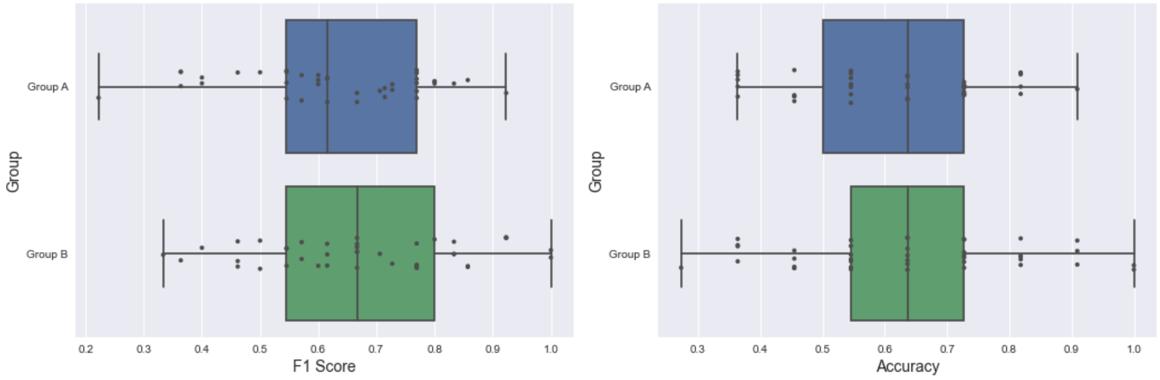


FIGURE 5

This boxplot figure reflects the F1 Score and accuracy of each subject in group A and group B. This graph is counted and sorted according to the TP, TN, FP and FN of each subject. We can clearly find out that the F1 Score of subjects in group B is better than the F1 Score of subjects group A, regardless of median, minimum, maximum or interquartile range. As for the accuracy, things are different. Since the interquartile range is also smaller in group A than in group B, the maximum and minimum are further from the median line. This means that, On the one metric of accuracy, there are more subjects in group B who perform better as well as worse than subjects in group A.

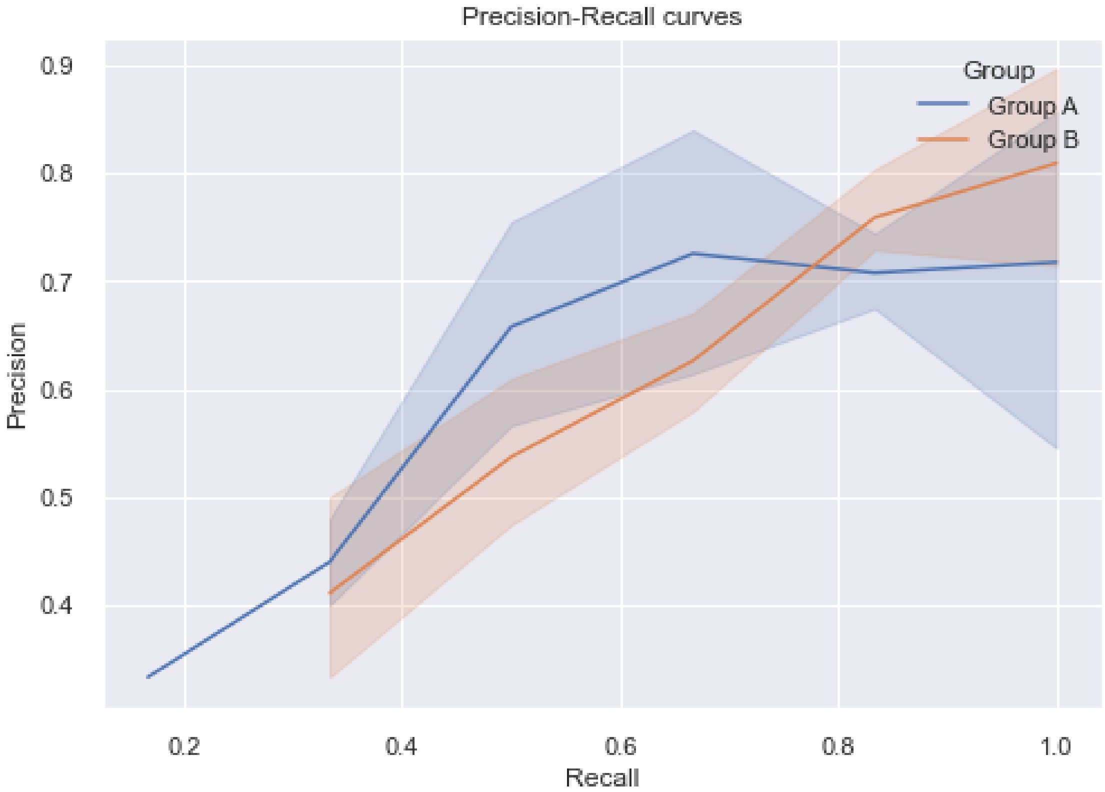


FIGURE 6

This lineplot figure shows the precision and recall scatter distribution for two groups. The area of the color reflects the number of repetitions of the point. The larger area indicates more repetitions of the same precision-recall value. The graph can reflect the overall distribution of the experimental results of the two groups. When the precision is the same, the recall of subjects in group B is significantly higher than that of subjects in group A. On the contrary, at the same recall, the precision of subjects in group B is significantly higher than that of subjects in group A. Besides, the intervention guarantees the lower bound of Group B on the precision and recall.

On the basis of the three figures presented, it is evident that the provision of DFD significantly impacts the performance of experiment, as the precision and recall of subjects in group B are consistently higher than those of group A. In contrast, the subjects in Group A had lower precision and recall compared to Group B. Moreover, the number of subjects in group B is much larger than that in group A when both precision and recall are high, indicating a more robust representation of the population. Furthermore, the results suggest that group B outperformed group A in terms of precision and recall, as indicated by the consistently higher F1 score. The findings indicate that providing DFD effectively improves the recall and F1 score of the subjects, confirming the importance of this approach in enhancing the performance of computer security control experiments. Overall, the results support the use of DFD as a key intervention to improve the accuracy and effectiveness of computer security control experiments.

6. Analysis Procedure

Describe the statistical tests used, including the explanation of which hypothesis the test is addressing. The characteristics of the presented data (in Section 5) must be in line with the assumptions (about the sample) that the test requires. Present your dependent / independent variables, and describe the steps you took to perform the analysis. Add graphics if appropriate. Describe the final results.

The central hypothesis of the experiment is DFD has significantly help in understanding the security threats of a system. This is because the intervention is designed and the control is implemented based on whether DFD is provided. In this section, we will use a statistical test to determine if the mean value of measurements of group A differs significantly from the mean value of measurements of group B. The t-test, as a typical statistical test is often used to compare the means of two groups of data and determine if the differences between them are statistically significant or simply due to chance. The t-test calculates a t-value, which is compared to a critical value to determine if the differences in the means are significant. If the t-value is greater than the critical value, then the difference between the means is significant, meaning that it is unlikely to have occurred by chance alone.

After finishing the experiment and receiving the results, first step is cleaning up the data. We did a rough comparison and statistics of the three tables. Although the number of subjects in each table, the number of people in the two groups is slightly different, after the deletion of invalid data, we use the understanding entry raw data as the most central table for data analysis, statistics and visualization. We clean the data based on two factors, one is the time to answer the question, and the other is whether the subject of the experiment clearly attends the training procedure. The reasons for the determination of these two elements have been analyzed in detail in the previous section. In short, this kind of dirty data will cause serious deviations in the experimental results.

6.1 Subject Self-evaluation

The next step is the subject background analysis. Too unequal experimental subject background will also cause serious biases in the experimental results. We organized according to Understanding_Threats file and visualized the relevant data. We already did this step before. The following figure serves as a supplement to TABLE 2 below, which shows the ratio of subjects with relevant backgrounds to the total subjects in each group.

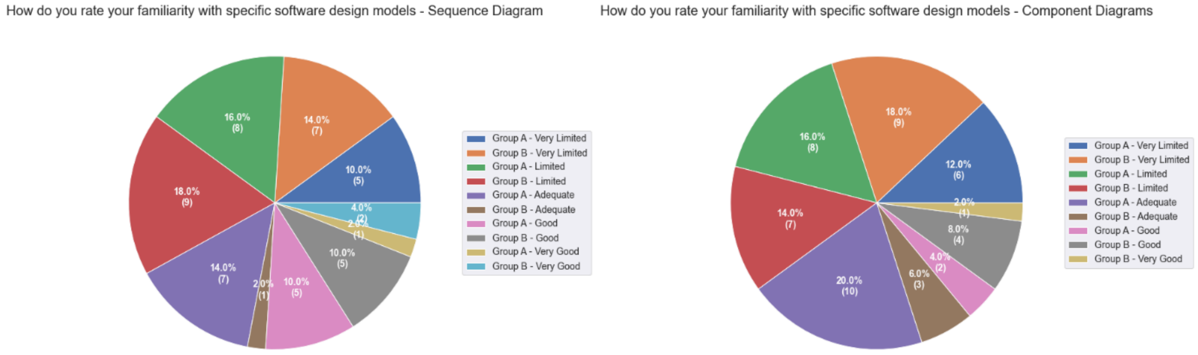
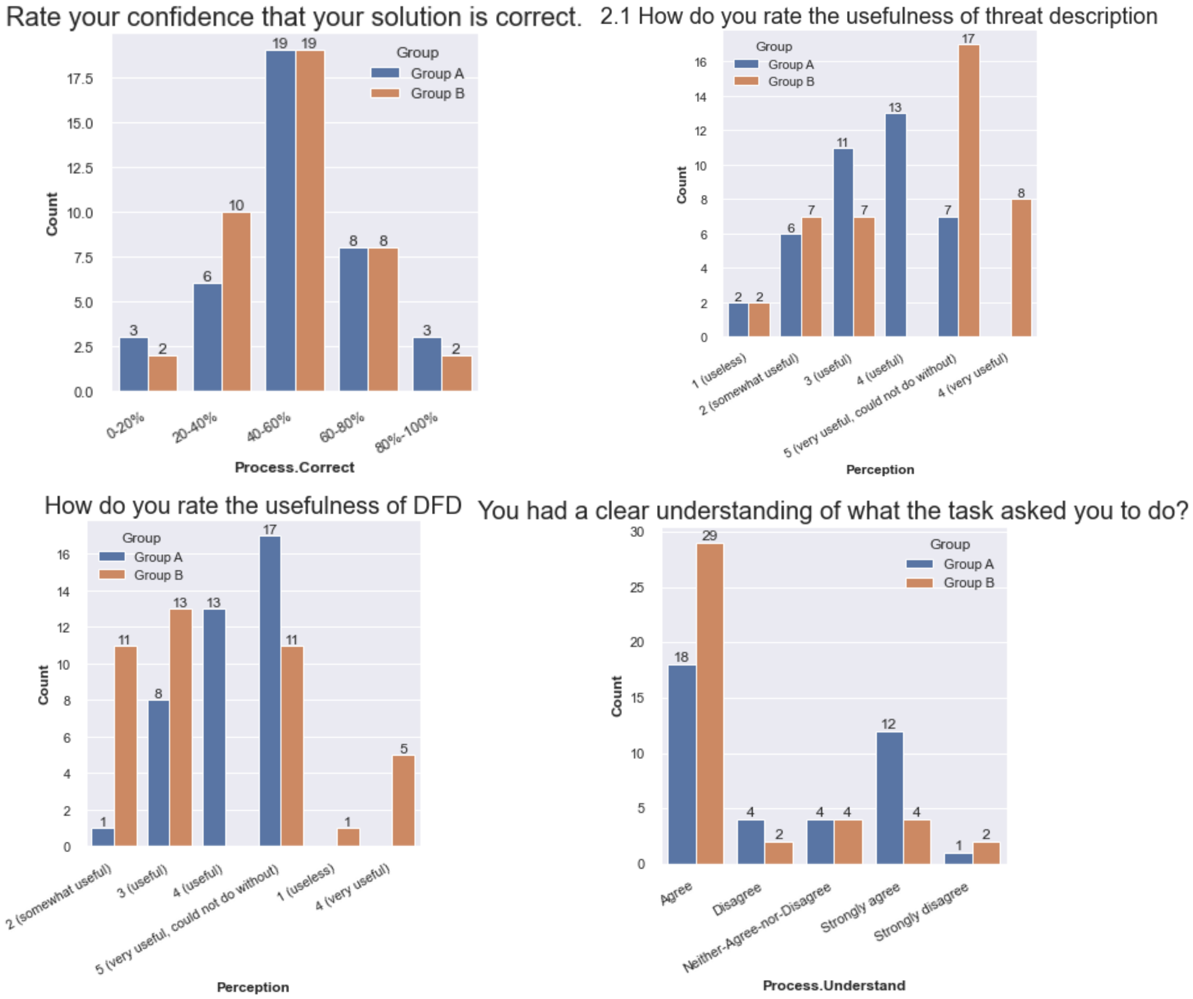


FIGURE 7

From the FIGURE 7 and TABLE 2, we can observe that, on the overall level, the backgrounds of group A and group B are not much different. This shows that the random assignment is relatively successful, and there is no situation in which the prior knowledge of one group is much higher than that of the other group in all aspect. However, according to the last three figures, we can clearly observe that there are many subjects in group B who have a very sufficient understanding of specific software design models, but there are no such subjects in group A. At the same time, because the number of subjects in group A is less than that in group B, the proportion of subjects in the two groups with very sufficient prior knowledge will further increase. This less-than-ideal situation may cause some disturbance to the experimental results. Based on the findings of this background check, we can reasonably assume that there may be subjects in group B have very high accuracy, precision and recall, and this kind of subjects may be significantly more than subjects in group A.



Rate your confidence that your solution is correct.

2.1 How do you rate the usefulness of threat description

How do you rate the usefulness of DFD

You had a clear understanding of what the task asked you to do?

FIGURE 9

After all the assumption and preparation, we analyse the data based on section 5 and get a conclusion. Based on the results in FIGURE 2, although the accuracy rates of the two groups are different for each single question, the subjects in group A are significantly better than the subjects in group B in half of the threat screening. Combined with in FIGURE 5 for overall analysis, we can conclude that the two groups have a slight difference in the accuracy of individual questions, but the overall accuracy is basically the same. This may be caused by DFD and can be reasonably infer that DFD should be of great help to the problem like MALICIOUS-CODE-GITHUB, but at the same time DFD may also mislead subjects in some questions and make subjects more radical. This could also explain the improved average recall for a population but not the improved precision. Anyway, the most valuable metrc, F1 Score, has indeed been significantly improved.

6.2 T-Test

The steps of using t-test on a computer security control experiment is as follows:

1. Identify the two groups you want to compare. In a controlled security experiment, these could be the group with the security control (intervention) and the group without the security control (control).
2. Decide on the outcome measure you want to compare between the two groups. For example, you might want to compare the accuracy or F1 score between the two groups.
3. Collect data on the outcome measure for both groups. Make sure to use the same method and data sources for both groups.
4. Calculate the mean and standard deviation of the outcome measure for each group.
5. Use a two-sample t-test to compare the means of the two groups. This test assumes that the data is normally distributed and that the two groups have equal variances. You can use the formula provided below to calculate the t-value.
6. Compare the calculated t-value to the critical t-value for your chosen level of significance and degrees of freedom. If the calculated t-value is greater than the critical t-value, the difference between the two groups is considered statistically significant.
7. Interpret the results and draw conclusions based on the statistical significance and the effect size of the difference between the two groups.

The t-test formula is as follows:

$$t = \frac{\bar{X}-\bar{Y}}{S_w\sqrt{\frac{1}{n}+\frac{1}{m}}}, \text{ where } S_w = \sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}$$

In this formula, $\mathrm{n} = 39$ represents the subject number in group A and $m = 41$ represent the subject number in group B. $\bar{X}$ represents the mean value of a metric for group A and $\bar{Y}$ represents the mean value of the same metric for group B. $S_{\bar{X}}$ represents the standard deviation of the same metric for group A.

Before the analysis, the null hypothesis need to be assumed, stating that there is no significant difference between the two groups in a specific metric. The significance level, alpha, was set at 0.05. In order to interpret the results of the t-test, we need to compare the obtained p-value with the significance level alpha.

Metric	t-value	p-value
Precision	-0.006695	0.994675
Recall	-1.703517	0.092452
Specificity	0.639382	0.524447
Overall accuracy	-0.732476	0.466073
F1 Score	-1.250000	0.215036

TABLE 5

The TABLE 5 shows t-test and p-value for five important metrics. Taking the F1 Score as an example, the obtained p-value is 0.215036, which is larger than the significance level of 0.05. This means that we fail to reject the null hypothesis and conclude that there is no significant difference between the means of the two groups at the 5% significance level. The t-value is -1.250000, which measures the size of the difference between the means of the two groups relative to the variability within each group. The negative sign indicates that the mean of group A is lower than the mean of group B. However, the absolute value of the t-value is not large enough to reject the null hypothesis. For the other metrics, none of the p-values were less than the significance level of 0.05, indicating insufficient evidence to reject the null hypothesis. Thus, we cannot conclude that the mean of these five metrics of group A is significantly greater than the mean of group B. Despite the inability to confirm a significant difference between the two groups based on the t-test, the negative t-value in precision, recall, overall accuracy, and F1 score suggest that the mean of Group A is lower than that of Group B for these four metrics. The results of this test also indirectly verified our above quantitative and qualitative analysis in Section 5.

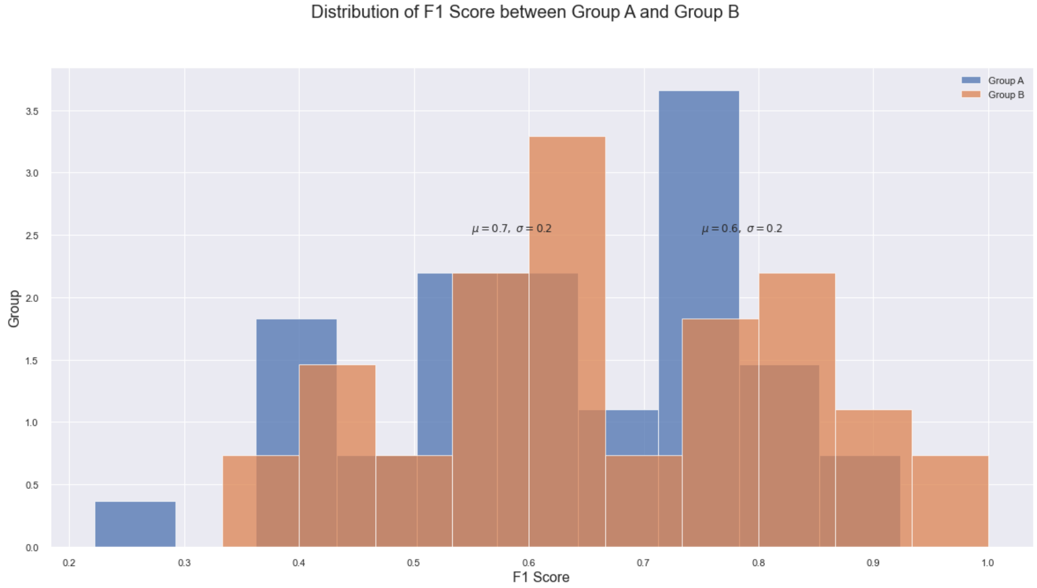


FIGURE 9

Summary of Experiments

End the report with your interpretation of the results and provide your main findings. Make sure that the main findings are in fact supported by the analysis described in Section 6 and explain how you derived those conclusions.
Eventually, you will have to exercise your judgement in determining whether the effect is actually practically significant or just statistically significant or insignificant from all perspectives.

Our experiment is a control security experiment. The subjects in two groups were trained with and without the Data Flow Diagram (DFD) as a control security mechanism. Based on all the introduction and analysis, we can get a conclusion that **whether DFD provided in the material have positive influence on the experiment results, but the influence is not significant enough**. We can find that the overall performance of group B is actually better than group A from the Section 5. This shows that the Data Flow Diagram has a useful and positive impact on understanding the security threats of Github. Conversely, there is no difference in the overall performance of the two groups in some metrics as precision, or even worse in metrics like specificity. It means that the intervention of DFD did not have a significant and positive effect on all the metrics in this experiment. The results of the t-test can also verify the above elaboration. Based on the analysis described in Section 6 and the results of the t-test, we can conclude that there is no significant difference between the means of the two groups for the five metrics measured (precision, recall, overall accuracy, F1 score) at the 5% significance level. However, we can still interpret the negative t-values for precision, recall, overall accuracy, and F1 score as evidence that the mean of group A is lower than the mean of group B for these metrics. This information may be useful for drawing conclusions about the effectiveness of the control mechanism being tested. Therefore, while the t-test does not provide strong evidence for a significant difference between the two groups in the five metrics measured, it does offer some insight into potential differences between the groups that may be worth further investigation. It is important to note that the effect of DFD may not be practically significant in our experiment. Thus, more comprehensive studies and analyses are needed to determine the practical impact of DFD on the overall security of the system.

In conclusion, our main finding is that the use of DFD as a control security mechanism can improve the subject's ability to find threats. However, the effect may not be practically significant, and further research is needed to fully understand the impact of DFD on the security of the system.