Automated Program Repair effectiveness

Experiment Reproducibility Report - E-3

Security Experiments and Measurements

Vrije Universiteit 2022

Student ID: 2722930
Student Name and Surname: Haohui Zhang
VU email: h17.zhang@student.vu.nl

Student ID: 2738163
Student Name and Surname: Behnam Bozorgi
VU email: b.bozorgi@student.vu.nl

Plagiarism statement

We hereby confirm that this assignment is our own work, is not copied from any other person's work (published or unpublished) and was not shared with any other person that could copy its contents.

Signatures:

Haohui Zhang

Behnam Bozorgi

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.

The aim of this experiment is two-fold. Its primary goal is to evaluate how well Automated Program Repairs (APR) tools help developers properly identify and fix vulnerabilities. Its second goal is to ascertain whether the knowledge that a security patch was generated by an allegedly specialized security tool influences the decision of human reviewers in evaluating the effectiveness of the patch. By examining these two goals, we hope to gain insight into the potential benefits and limitations of using Automated Program Repairs tools in the context of software security, as well as the role of human judgment and perception in evaluating such tools.

Totally seven vulnerabilities are needed to be distinguished and seven tools are taken into the consideration, separately Arja, Card, jGen, jKali, RSRepair, TBar and SeqTrans. For the first part, the intervention should be the actual APR tool, as the intervention group is provided the actual APR tool and the control group is provided the fake APR tool or no APR tool. However, with regard to the fact of one single group, all students received training on using a plugin for selecting and visualizing the vulnerabilities and were provided with the identical plugin in the experiment. Specifically, all subjects were given a plug-in constituted by 5 APR tools which allows them to see how the patch would be applied in the code and to navigate through the different patches. Therefore, there is no control group in the first part.

As for the second part, the intervention is **the additional information about the type of the APR tool (generic vs security)**. The second part of the experiment involved presenting each subject with additional information regarding the recommended patch for each vulnerability. Specifically, this information included the proposed patch suggested by a security-specific tool. However, the type of tool used was varied between security-specific and generic, which separately provide the correct information and wrong information. The aim of this part of the experiment was to determine whether providing this information would influence the decision of the developers, regardless of whether the information provided was correct or incorrect. Therefore, for each vulnerability, we can divide all the subjects into two groups, the intervention group is provided with the actual security-specific tool which is the SeqTrans, while the control group is provided with the generic tool. Note that, the recommended patches provided by SeqTrans are not always the correct patch.

2. Metric(s) 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.

<write your text here>

The purpose of this section is to determine if the intervention succeeds or not. To measure the success of the intervention, firstly we need to preprocess the data and determine the criteria for correctness (true positive, false positive, true negative and wrong negative) and then use the evaluation metrics to measure the success of the intervention. The metrics we decided in this report are based on the two goals and the two research questions respectively.

RQ1: *The actual effectiveness of APR tools in terms of finding vulnerabilities and helping developers to make decisions.*
Goal 1: To evaluate how well APR tools help developers properly identify and fix vulnerabilities.
Subject-wise metrics:

1. Number of correct patches (TP) identified as correct by the subject
2. Number of incorrect patches (TN) identified as incorrect by the subject
3. Precision, Recall, F1 score and Accuracy of each subject (Precision=TP/(TP+FP); Recall=TP/(TP+FN); Accuracy=TP+TN/(TP+TN+FP+FN); F1=2*(Precision*Recall)/(Precision + Recall))

Vulnerability-wise metrics:

1. Number of subjects who correctly identify one specific path
2. Number of subjects who incorrectly identify one specific path
3. Precision, Recall, F1 score and Accuracy of each vulnerability

General metrics:

1. Mean, Standard deviation, etc. of subject-wise metrics
2. Mean, Standard deviation, etc. of vulnerability-wise metrics

Goal 2: To ascertain whether knowing that a patch was produced by an APR tool does change the decision of human reviewers.
Group-wise metrics (as well as vulnerability-wise):

1. Number of switches and insists after the security information is revealed
2. Influence ratio

General metrics:

1. Mean, Standard deviation of influence ratio

RQ2: *The perceived usefulness of the APR tools as collected by participants' questionnaires for each vulnerability.*
We quantify this research question by assigning "0" to "Insist" and "1" to "Change" to see the perception or confidence of the subject of the usability of the various tools.

The definition of influence rate is as follows:
Based on the FIGURE 1, for each vulnerability and for each subject, he/she has three choices: Insist, Follow or Change, which will result in his/her final answer correct, partial correct or wrong. To quantify the influence of APR tools on subjects' answers, we assign "2" to "Correct", "1" to "PartialCorrect" and "0" to "Wrong". We conduct the same operation for their original answers. We then calculate the difference of the score on final answers and original answers for each subject, and we dub it as the influential ratio. Furthermore, we can calculate the mean score among one group.

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.). ***Explicitly mention the data on the past study and how this could impact your results.***

Participants of the experiment are master students from VU Amsterdam and all of them are computer science backgrounds. There are totally 80 subjects attending the experiment while eight of them don't consent to use their data.

The main step of the experimental process is the execution of the intervention. The replicable process is specifically as follows:

1. ATTENDED training session

All participants have received the identical training material on using a plugin for selecting and visualising the vulnerabilities, as well as how to use the plug-in to identify vulnerabilities. There is only one single group. Furthermore, the teacher maintains control over both groups by ensuring that each group was given the same amount of reading time.

2. PARTICIPATED in an experiment

The second step is conducting intervention. After reading and watching all the material, the experiment is conducted. Firstly, the relevant experience on security testing of the subjects is investigated. Then, participants will receive a number of projects to evaluate and are required to analyze the vulnerability and the corresponding patches suggested. There are 4 or 5 suggested patches per vulnerability which are all generated by the APR tools. Only one is the security tool and else are generic tools. For every suggested patch, the participants are requested to evaluate it by grading it CORRECT, PARTIALLY CORRECT, PARTIALLY WRONG or WRONG. After evaluating all the patches for each vulnerability, participants will decide to pick one patch (must only pick one) to pass for further code review as well as to suggest one patch to adopt (also no patch might be adopted).

3. ANSWERED question in the Qualtrics survey

After identifying the vulnerabilities and picking one patch for each vulnerability, participants are asked whether they will confirm the chosen patch when being provided additional information that one of the patches was provided by a security tool. The intervention for the second part is conducted as all participants are randomly assigned to a group, the additional information (real security patch) for group A is suggested by the security tools, and additional information (bogus security patch) for group B is suggested by the generic tools. The participants then need to decide to change their original selection by following the suggestion provided by the tool or insist on their original selection. Note that changing their original selection but not to the suggested patch is also possible here. Although we don’t know the intuition behind this, we can use these data as a noise. Eventually, all participants are requested to fill out a survey, which was about evaluation on the whole task but nothing on their perceptual feelings on APR tools. This evaluation only reflects the subjects’ knowledge of the material.

4. DETERMINED the ground truth

All the answers are recorded, the related time is recorded, and all the relevant information is aggregated. All data are aggregated into two Excel files according to the training and experiment procedures, combining with two Excel files with last year data. We can use two types of validation procedures to find the true positives based on the ground truth: automatic or manual. An automatic validation procedure would use an algorithm or code program to validate the results, while a manual validation procedure would involve human input to validate the results. We also need to determine the error rate of this algorithm by two independent assessors with an agreement or consensus meeting.

The below two processes are not exactly part of the experiment process, but we put them here just for the sake of clarification.

5. PREPROCESSED the collected date

We performed data cleaning on the data in four excel, specifically deleting the dirty data based on the duration time, whether they consent to the usage of their data and whether they took the training procedure. After that, we matched the participants' answers with the ground truth, and identified our own criteria to get true positives (more details will be elaborated in Section 6). Then, we proposed various statistics measures for evaluation, plotted the data for easier comparison and elaborated the results quantitatively and qualitatively.

6. CONCLUDED the results and experiment

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. We also conducted the same procedure for last year's data but only for a simple comparison as the absence of a lot of important relevant information. Note that we didn't combine the last year data and this year data together to conduct the statistical test because of the ambiguity of the ground truth for the last year's data. Therefore, the past data could not impact the final results. 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.

Explicitly mention if the data on the past could impact your results.

The key limitations of this experiment design are the sample size, the subjects' background and the randomization. Due to the small sample size, the statistical power of the experiment is limited. A good solution would be to combine the past data with this year's data to expand the sample size. However, due to the last year code files being not provided and the absence of some relevant data, after careful deliberation, we decided to not combine two years data together and only use the past data as a control data. Secondly, the experiment does not account for the confounding variables that may affect the ability of the subjects to identify vulnerable lines, such as their level of expertise in the field, level of expertise on java or attention during the training. Randomization is not a big problem in this experiment as the participants are not grouped in part one, and in part two, all the participants are randomly divided into two groups for each vulnerability.

The key criticalities of this experiment design are the intervention implementation. 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, we can not obtain any valid conclusion from part two of this experiment. The determination of criteria on defining correctness (true positive) is also very important, if we use a too strict criteria, the true positives and true negatives for all subjects may be too less and the distribution may be highly biassed. We are not able to generate any effective conclusions from highly biassed results. If the constraint is too optimistic, the persuasiveness of the results and scalability of the experiment will be too weak.

Ground Truth Decision

5. Definition

You have collected from the subjects some experimental artefacts and you should specify what makes the artefact measurable (for example correct or incorrect threats) or not qualified for evaluation of the success metrics (for example missing values or fields). If you have multiple measures of success (for example finding two types of vulnerabilities) the definition should be given for each measure.

CVE	**TYPE**	**Arja**	**Card**	**jGen**	**jKali**	**RSRepair**	**TBar**	**SeqTrans**
CVE-2018-1192	Information Disclosure	C	P	PC			C	P
CVE-2019-10173	Serialization	PC	P	C	PC			P
CVE-2016-9878	Path		P	P	P		P	P
CVE-2018-1324	DoS	C			C	C	P	P
CVE-2013-4378	XSS	PC				PC	PC	C
CVE-2018-17202	DoS	PC	PC		PC	PC		P
CVE-2018-1000864	DoS			P	P	P	P	P

TABLE 1

The above table presents the ground truth provided by the teacher. However, there are some distinctions between the ground truth and the experiment data. The ground truth of each APR tool on each CVE only has three cases, separately CORRECT, PARTIAL CORRECT and WRONG. However, during the experiment procedure, for all CVE, all subjects had four choices to select, respectively CORRECT, PARTIAL CORRECT, PARTIAL WRONG and WRONG. Because of the existence of the PARTIAL CORRECT and PARTIAL WRONG, we caution that leveraging the data points of all subjects to draw effective conclusions necessitates criteria to determine the true positives and true negatives. The universal similarity of all criteria is that successfully labeling the CORRECT patch is defined as the true positive, and successfully labeling the WRONG patch is defined as the true negative. The restriction ranges from strong to weak in the below three criteria.

1. Successfully labeling the PARTIAL CORRECT patch is defined as the true negative, and labeling the patch as PARTIAL WRONG is defined as the false negative.
2. Successfully labeling the PARTIAL CORRECT patch is defined as the true negative, and labeling the patch as PARTIAL WRONG is defined as the true negative.
3. Successfully labeling the PARTIAL CORRECT patch is defined as the true positive, and labeling the patch as PARTIAL WRONG is defined as the true negative.

After the criteria are selected, we calculate the aforementioned metrics which are elaborated in Section 2 to evaluate the success.

For the part 2, the defining of true positive is more complicated. As we randomly separate the subjects into groups for each vulnerability, therefore, there are 7 different intervention and control groups. FIGURE 1 shows how the procedure progressed in part 2. In this part, we follow the second criteria defined above, as the first one is too strict, and the details of the process will be discussed in the next section. For each vulnerability, each subject will choose to “Insist” his/her original decision, or “Follow” the suggestion provided by the (generic or security) tools, or else “Change” his/her original decision to a new one, which is also not the suggestion. These three actions further separate the subjects into three different subgroups, and after that, we obtain each subject’s final decision. If their final decisions are the CORRECT or PARTIAL CORRECT patches, we assigned these data points as the true positives. If their final decisions are the WRONG patches, we assigned these data points as the false positives. Note that, in this case, there are no true negatives and false negatives, only correct (true positives) or incorrect (false positives). Note that the vulnerabilities “CVE-2018-1000864” and “CVE-2016-9878” are not qualified for evaluation of the success metrics, as all patches for them are wrong. Therefore, we decide to exclude these two vulnerabilities and mainly consider the else 5. Except for count or mean of the correctness for each subgroup, a new metric is defined in Section 2. We dub this metric as influential rate and use it to conduct further analysis for part 2.

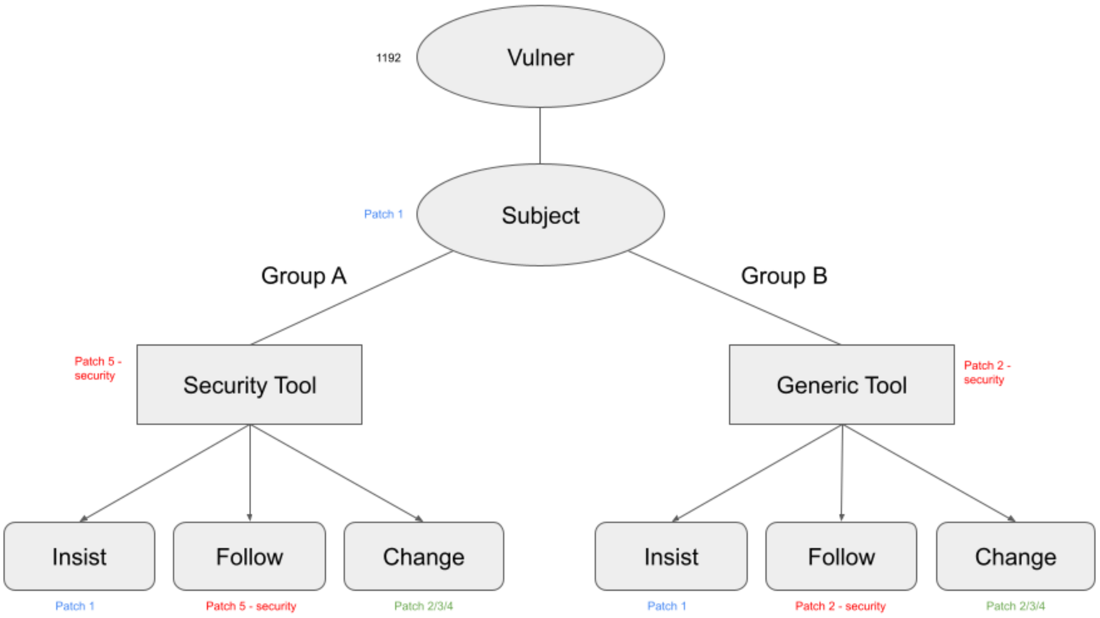

FIGURE 1

6. Process and Guidelines

You might give both an automatic or a manual validation procedure. If you give an automatic procedure you should specify the algorithm and what are the possible errors that this might introduce in the measure of success. For a manual procedure you should start by assigning artefacts to different group members, start a preliminary evaluation, come to common scoring guidelines (reported), and then continue scoring.

After defining the criteria of the correctness evaluation, we need to apply a ground truth generator to find the true positives for each subject. We applied an automatic procedure, which is a python script, to evaluate subjects' artefacts. We didn't apply the manual one as the data points are too much, and manually checking the ground truth one by one for two years data is basically too inefficient. The automatic procedure contains three steps, respectively artefacts preprocessing, criteria selection and correctness determination.
Artefacts preprocessing: For part 1, the data points are easy to sort out. We extracted all subjects' evaluations on patches per CVE separately based on the column number. Right now, we have the evaluation of each subject on each patch for each vulnerability, and we put it into a big dictionary. As for the data points of part 2, we only extract subjects' original answers and their final answers for each vulnerability.

Criteria selection: After the preprocessing, we first extracted the ground truth from the qualtrics-mapping.xlsx and transferred the ground truth into the wanted data type (dict). We then mapped the ground truth dictionary with the subject dictionary to obtain the true positives, etc. We implemented the second and third criteria separately. The reason why we abandon the first one is because of the highly biased results.

Correctness determination: For part 1, we already obtained the correctness, and we can use the output to calculate the precision, recall, accuracy for each subject. As for part 2, we firstly determined the correctness of each subject's original decisions based on the ground truth. Then, given that the subjects' final answers have two versions with "- security" or without the hyphen and the knowledge of only SeqTrans is the security tool, we successfully classified our data in two groups (Group A and Group B). Group A is the intervention group which is provided the real information given by SeqTrans (second column in the experiment excel), while Group B is the control group which is provided the bogus information given by a generic tool (first column in the experiment excel). Then we drew a comparison between subjects' original answers and final answers to determine whether they insist on their opinions or follow the suggestions, or change their minds. An example is presented in FIGURE 1, and we think the further explanation of this step is trivial. Finally, we classified all subjects' data based on their actions and their group, and checked the correctness of their original answers and their final answers by mapping their results with the ground truth dictionary. We then assigned a score for the correctness and calculated the influential rate for each subject, and compared the mean of influential rate among the dimension of group. Note that, as one subject can both "Insist" and "Follow" based on the above definition, we decided to classify these overlapping data points into "Insist" class. However, this may introduce some bias.

There may be some possible errors regarding the correctness determination step for part 2. All further processing and analysis are based on the initial assumption that **data points containing "- security" are the suggestions provided by the (generic or security) tools and all the subjects in one group will receive the same suggestions**. However, due to the lack of information provided in the data, it is difficult to compromise the validity of the assumption. An incorrect initial assumption can result in making the whole analysis invalid. However, after we fully investigated the data, we think our initial assumption is correct. Quantifying the correctness may also cause some errors, but our algorithm effectively avoids this problem, and this can be verified in the next section.

7. Validation and Error Correction

If you do an automatic analysis you should select a random subset of artefact which you manually validate to determine the error rate of the automatic process.

We adopted a manual validation procedure for our algorithm. Firstly, we randomly sampled 10 artefacts in a random sequence and assigned them to different group members, which are two independent humans. All the artefacts were unmarked. We applied the human assessors with conflict resolution method as in FIGURE 2. The preliminary evaluation would involve each group member individually assessing the artefacts they were assigned and providing a preliminary score. The score is defined as a binary (whether "0" or "1") for both parts in this experiment. Then, my teammate and I worked together to come up with common scoring guidelines. During the discussion, we came to a conclusion that the explanation texts are really an important reference and we decided to apply the second criteria as the correctness criteria in this procedure. Once the common scoring guidelines are established, the group members would continue scoring the remaining artefacts according to these guidelines. This approach is intended to minimize the potential for bias or inconsistency in the manual validation process, since scoring is independent. The correctness assignment of our aforementioned automatic analysis is all correct. With regard to the low error rate, we didn't apply the error correction for the misaligned results. As for part 2, we randomly selected two vulnerabilities and manually checked all the data points containing "-security" for each column whether having the identical value or not. We didn't follow the procedure in FIGURE 2 in this part. The result is they all have the same value for each column. We also checked whether the same subject will be assigned to the same group for each CVE. We found that the assumption is not established. Therefore, we confirmed our assumption.

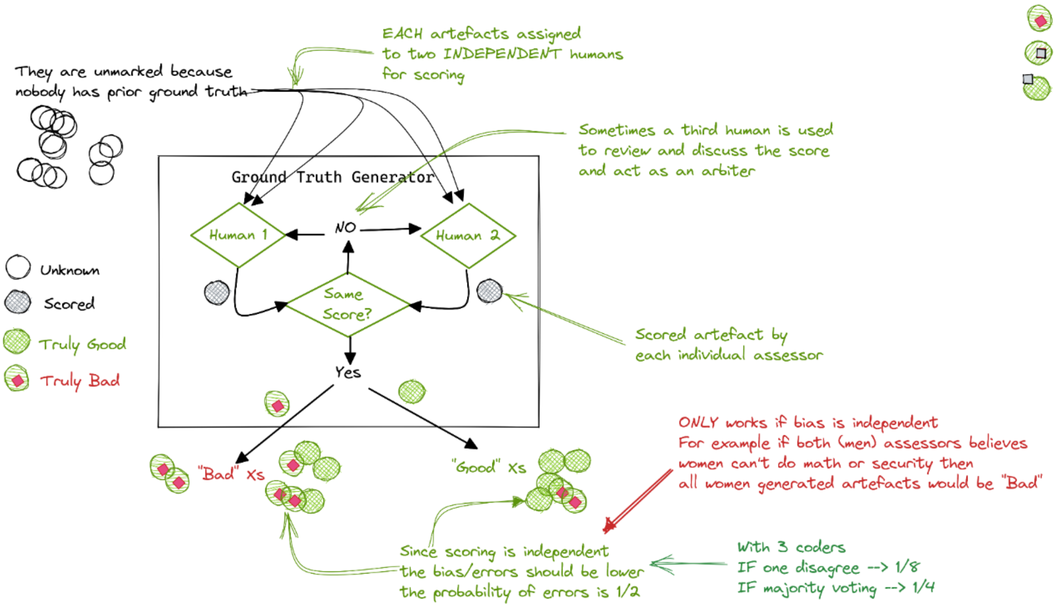
EACH artefacts assigned
to two INDEPENDENT humans
for scoring
They are unmarked because
nobody has prior ground truth
Sometimes a third human is used
to review and discuss the score
and act as an arbiter
Ground Truth Generator
Human 1
NO
Human 2
Same
Score?
Yes
Unknown
Scored
Truly Good
Truly Bad
Scored artefact by
each individual assessor
ONLY works if bias is independent
For example if both (men) assessors believes
women can't do math or security then
all women generated artefacts would be "Bad"
"Bad" Xs
"Good" Xs
Since scoring is independent
the bias/errors should be lower
the probability of errors is 1/2
With 3 coders
IF one disagree --> 1/8
IF majority voting --> 1/4

FIGURE 2

Analysis Description

8. 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. ***Explicitly mention if the data from the past has been merged or is analyzed separately.***

After determining the ground truth from the file, we first preprocessed the data and filtered out the invalid data, e.g. replicate data. Then, we visualized the duration time to find the outliers. Luckily, according to FIGURE 3, nearly all subjects in two years finished in the requested time, and the timeout is not exceeded too far. We can observe that this year subjects spent more time on the experiment in average than last year. Following the preprocessing stage, we established three specific criteria and subsequently implemented an automated procedure to ascertain the true positives and true negatives for each subject. These values were then utilized for further analysis of the data, using basic descriptive statistical measures. In this section, we conduct interpretation and analysis based on two research questions for 2 goals. When analyzing RQ1, we also conduct in-depth quantitative and qualitative analysis based on different aggregation methods (subject-wise, vulnerability-wise or general). Note that we didn't merge the two years data together because of the ambiguity of the ground truth for the last year data. We just analyzed them separately and compared the results.

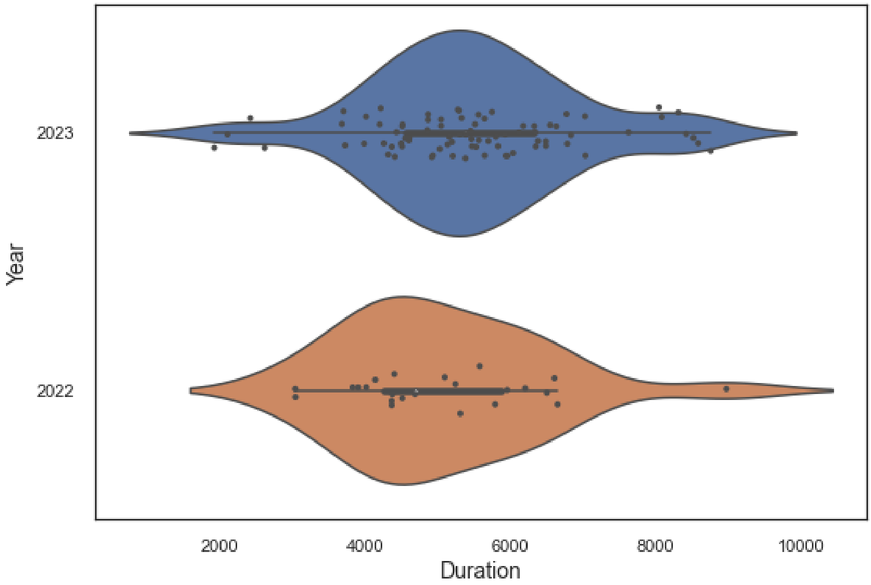

FIGURE 3

Goal 1: Vulnerability-wise metrics

The TABLE 2 below presents the number of subjects who correctly identify one specific vulnerability (true positives and true negatives) and the number of subjects who incorrectly identify one specific vulnerability (false positives and false negatives) by applying Criteria 3. The TABLE 3 is formatted in the same way as TABLE 2 but by applying Criteria 2.

CVE	TP	FP	TN	FN
CVE-2018-1192	63	180	103	59
CVE-2019-10173	77	166	114	48
CVE-2016-9878	0	0	280	125
CVE-2018-1324	110	133	124	38
CVE-2013-4378	135	189	0	0
CVE-2018-17202	64	260	35	46
CVE-2018-1000864	0	0	276	129

TABLE 2

CVE	TP	FP	TN	FN
CVE-2018-1192	22	140	110	124
CVE-2019-10173	18	63	140	184
CVE-2016-9878	0	0	280	125
CVE-2018-1324	53	190	124	38
CVE-2013-4378	33	48	49	194
CVE-2018-17202	0	0	85	320
CVE-2018-1000864	0	0	276	129

TABLE 3

By comparing the above two tables, we can clearly observe that the Criteria 2 will result in more negatives than positives. You can see that the value of the true positives are significantly shrinked. We also notice that CVE-2016-9878 and CVE-2018-1000864 only have negatives as the ground truth of all patches for them are all WRONG. Therefore, there is no need to compare the precision or recall for them. In the following section, we decide to use the Criteria 2, as the values of positives and negatives are even-distributed. From the TABLE 2, we can see that a lot of subjects perform pretty well on evaluating the CVE-2019-10173, CVE-2018-1324, especially for the CVE-2018-1324, more than half of the subjects' answers are correct. But for CVE-2018-17202, the performance is pretty bad. Based on our analysis, this phenomenon may be caused by the different difficulty of evaluating different vulnerabilities.

The FIGURE 4 shows the accuracy, precision, recall and f1 score for each vulnerability by using the positive and negative values in FIGURE 2.

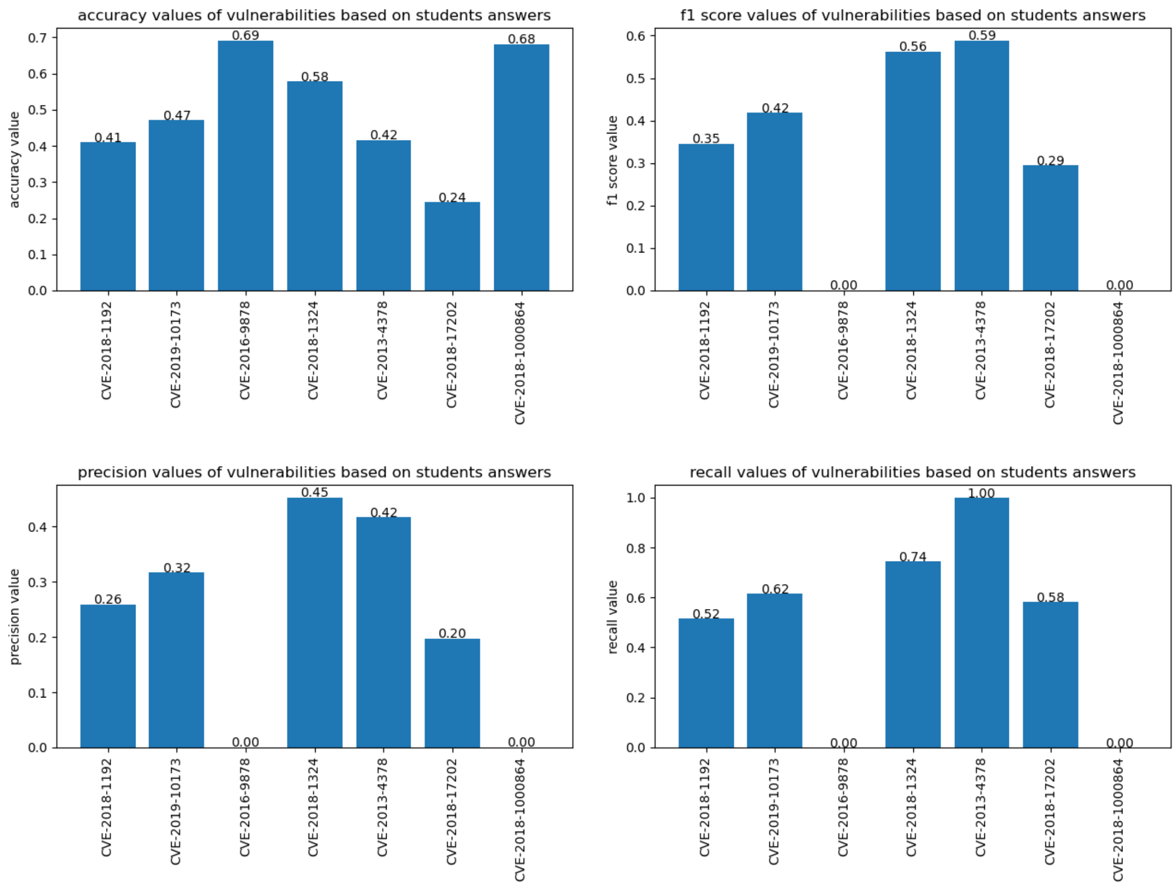

FIGURE 4

Drawing insights from the upper left bar plot depicted in FIGURE 4, it becomes evident that participants exhibit outstanding performance on CVE-2016-9878 and CVE-2018-1000864, despite their invalidity in the other metrics. Apart from these two, participants' performance on CVE-2019-10173 and CVE-2018-1324 corroborates our previous analysis, exhibiting satisfactory results. In terms of the F1 score, CVE-2018-1324 and CVE-2013-4378 demonstrate the highest performances. Notably, the excellent performance on CVE-2013-4378 can be attributed to the application of Criteria 2, leading to only positive values being available, and thus, the recall value is 1. Except for CVE-2013-4378, subjects have the best performance on CVE-2018-1324 regardless of accuracy, precision or recall.

Goal 1: Subject-wise metrics

After conducting a comparison between the vulnerability-wise metrics using Criteria 2 and Criteria 3, we have concluded that applying only Criteria 2 will be appropriate for our subsequent analysis.

Totally, there are 34 patches to evaluate for each subject. After the preprocessing, we obtain the true positives, true negatives, false positives and false negatives for each subject. In the upper box plot of FIGURE 5, we plot all subjects' amount of correctness, which can be gotten by adding true positives and true negatives. It is the sum of the number of correct patches (TP) identified as correct by the subject and the number of incorrect patches (TN) identified as incorrect by the subject. In the lower box plot of FIGURE 5, we plot all subjects' accuracy. Each black point in the figure represents a single subject. We can easily infer that most subjects do not perform really well regardless of the mean, median, minimum, maximum or interquartile range.Only a handful of participants demonstrate an accuracy as high as 0.64, which is clearly not good enough.

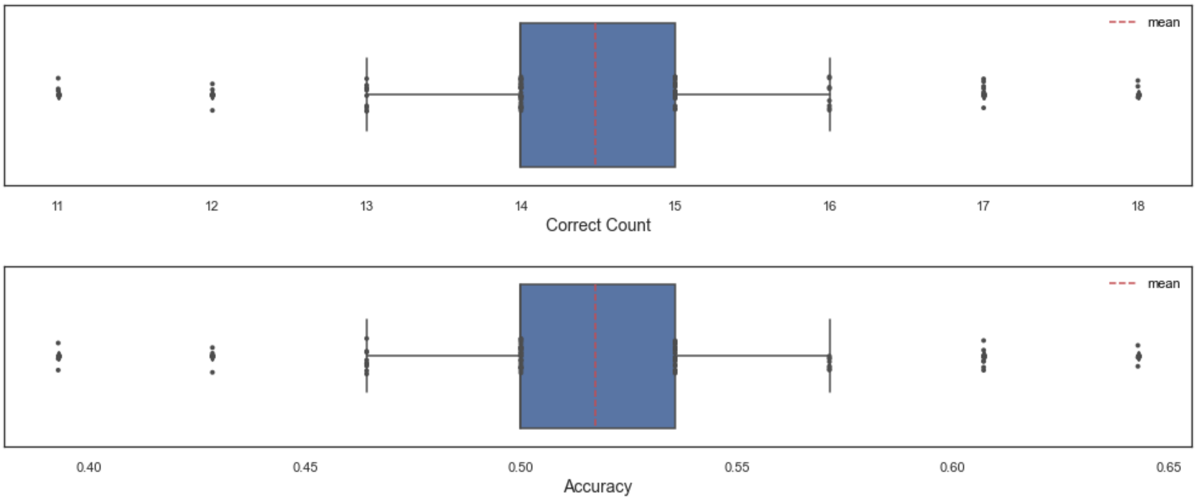

FIGURE 5

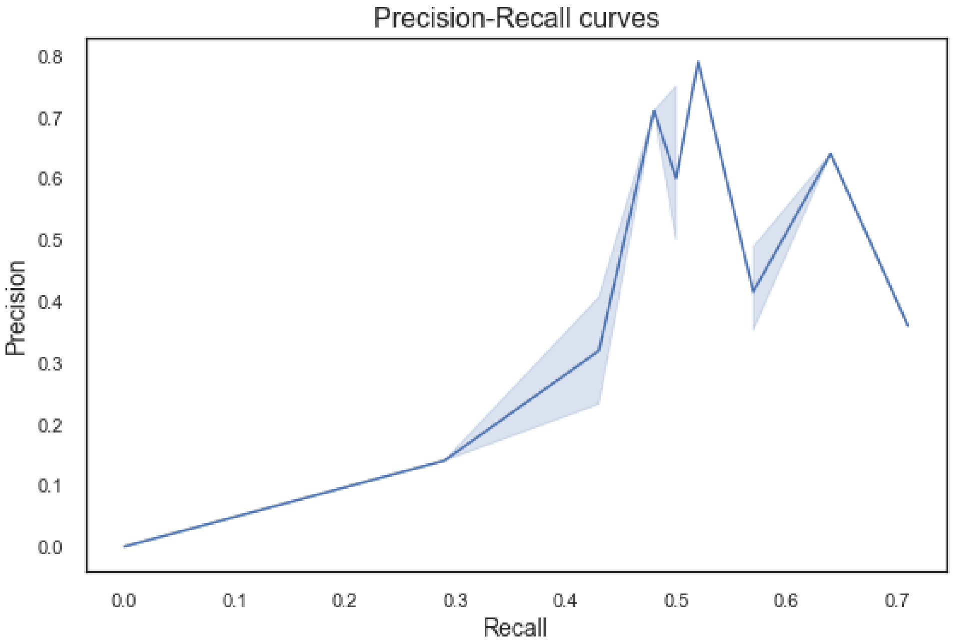

FIGURE 6

FIGURE 6 shows the precision and recall scatter distribution for all subjects. 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 partially reflect the overall distribution of the experimental results of the subjects, and the correlation between recall and precision. It is apparent that the precision and recall are linearly correlated for when the recall is lower than 0.5. But when the recall is higher than the 0.48, oscillations show up and there is no significant correlation between them. This indicates that attaining both high precision and recall is really difficult in this experiment. Normally, having a higher precision will damage the performance in recall. Conversely, the opposite holds true as well.

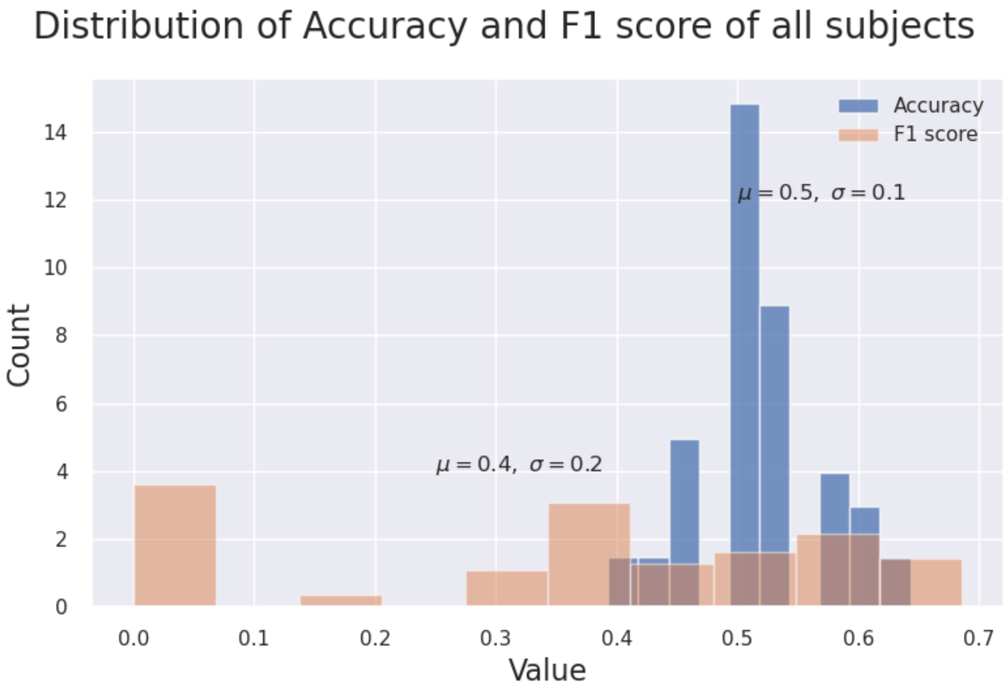

FIGURE 7

FIGURE 7 is the histogram on the accuracy and f1 score for all subjects. It reflects the overall distribution of the experimental results. We notice that the subject-wise accuracy precisely follows the normal distribution, but the subject-wise f1 score does not.

Goal 1: General metrics

In this part, we calculated the mean, median, standard deviation, geometric mean, etc. for subject-wise metrics and vulnerability-wise metrics separately. We classified them as the general metrics. Accuracy represents the overall ability to detect both correct and incorrect patches, whereas precision measures the ability to detect only correct patches with a high degree of accuracy. Recall, on the other hand, reflects the degree of randomness in a subject's decision-making process. Therefore, these metrics provide valuable insight into the efficacy of a subject’s performance. It is worth noting that accuracy, precision, and recall are not independent metrics and are often interrelated, meaning that an improvement in one may lead to a decline in the others. Therefore, it is crucial to strike a balance between these metrics to evaluate the optimal performance.

FIGURE 8 depicts a box plot of the precision, recall and accuracy among all subjects achieved for all patches when applying Criteria 2. It is apparent that subjects don’t detect much in general as the mean (red point) and median (orange line) of precision is really low. However, a few subjects have a really high precision or a really low one. Based on the low mean of recall, we can infer that subjects' detections are highly random. We can also infer that no one’s answer in this experiment is not correct.

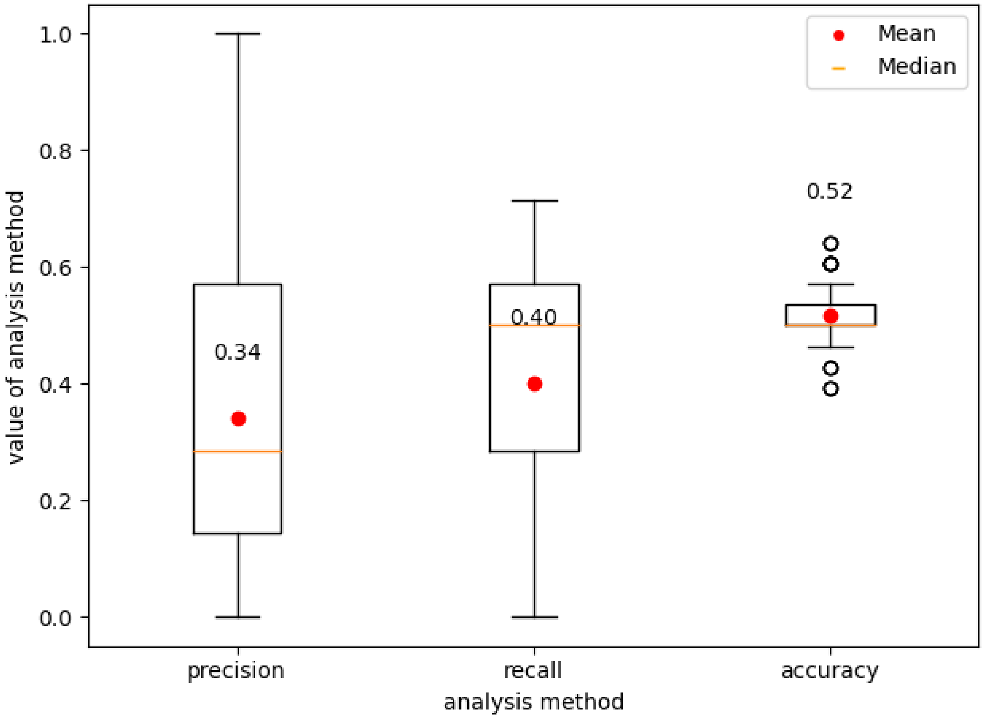

FIGURE 8

	Mean	Geometric mean	Standard deviation
Vulnerability-wise Accuracy	0.499029	0.473792	0.161646
Subject-wise Accuracy	0.517196	0.514414	0.053901

TABLE 4

Our results table (TABLE 4) is formatted so that each row corresponds to a particular perspective evaluation metric with columns giving the scores of different operations. We have specifically focused on computing accuracy as it is considered to be the most comprehensive and effective metric for evaluating a subject's performance across all cases. In the context of our experiment, the accuracy computed at the vulnerability-level serves as the level of difficulty for each vulnerability in terms of detecting both correct and incorrect patches accurately for all subjects. Conversely, the accuracy computed at the subject-level provides insights into the individual performance of each subject in detecting both correct and incorrect patches accurately. The mean of these two accuracies show the overall performance. From TABLE 4, we can see that the accuracy doesn’t vary a lot on the subject-level. However, the vulnerability-wise accuracy, as evidenced by the high standard deviation, exhibits considerable variability, indicating a wide range of difficulty levels associated with each vulnerability. Notably, some difficult questions clearly pull the mean and geometric mean down compared with the subject-wise accuracy.

Goal 2: Group-wise metrics

Number of switches (follows, changes) and insists after the security information is revealed in TABLE 5, and number of CORRECT, PARTIAL CORRECT and WRONG final answers conducted by all subjects is revealed in TABLE 6. Note that, the existence of “Change” does not logically hold, and therefore, we regard them as the noise in the later analysis. For sake of the clarification, “Security tool” represents Group A which is the intervention group, while “Generic tool” represents Group B which is the control group.

	Security tool			Generic tool		
	Insist	Follow	Change	Insist	Follow	Change
CVE-2018-1192	29	12	0	28	11	1
CVE-2019-10173	25	12	2	22	19	1
CVE-2016-9878	24	15	1	25	16	0
CVE-2018-1324	31	8	2	27	11	2
CVE-2013-4378	25	16	0	29	10	1
CVE-2018-17202	21	17	3	33	7	0
CVE-2018-1000864	28	11	2	31	9	0

TABLE 5

TABLE 5 presents some data about subjects’ final choice, which we will discuss. Firstly, we need to define each term in the table. “Insist” means that the subject's patch choice was the same as the patch which was indicated as a security tool patch by the survey. “Follow” means that the subject patch choice was different, and the subject chose to change his choice to the one that the survey mentioned. “Change” means that the subject changed his patch choice, but to a patch that is not the one that was presented by the survey. Also, by “Security tool” we mean that the patch which was presented to the subject was “real security tool” (security group) and “Generic tool” means that the patch which was presented to the subject was actually a bogus tool (generic group).

TABLE 5 shows that most subjects tend to stick to their choice, indicating a strong adherence to their initial selection. Some subjects, however, appeared to rely on the "security specific tool" option without much deliberation. And only very few subjects that just tend to just randomly change their choice, which we believe were just missclicks. This shows that most subjects were confident with their answer (Insist group) regardless of whether it matched the patch described as the "security specific tool" or not.

	Security tool			Generic tool		
	Correct	PC	Wrong	Correct	PC	Wrong
CVE-2018-1192	11	4	26	10	2	28
CVE-2019-10173	10	6	23	11	2	29
CVE-2016-9878	0	0	40	0	0	41
CVE-2018-1324	38	0	23	18	0	23
CVE-2013-4378	16	25	0	32	8	0
CVE-2018-17202	0	27	14	0	8	32
CVE-2018-1000864	0	0	41	0	0	40

TABLE 6

TABLE 6 shows how subjects performed in each vulnerability on their final answers (after they had a chance of changing their mind). We can see that in CVE-2016-9878, CVE-2018-17202 and CVE-2018-1000864, there is no correct answer because no patches were correct in these vulnerabilities in ground truth. It looks like most subjects had no problem on detecting CVE-2018-1324 correctly in security group and CVE-2013-4378 in generic group.

Goal 2: Influential rate

To quantify the influence of APR tools on subjects' answers, we assign “2” to “Correct”, “1” to “PartialCorrect” and “0” to “Wrong” based on TABLE 6. We can end up scoring the subjects based on patches they finally selected. We conduct the operation for their original answers (Original score) or their final answers (Final score). Likewise, we calculated the difference of the Final score and Original score (Influential_rate = Final_score - Original_score) for each subject and obtain the Influential rate. Besides, we decided to exclude the data points of CVE-2016-9878 and CVE-2018-1000864 in this part, as all patches' ground truth are labeled as WRONG.

FIGURE 9 depicts the mean of all subjects' final scores on the vulnerability-level, and the average of these points is shown. It can be inferred that vulnerability CVE-2013-4378 is far easiest for subjects to detect the correct patch, comparing with the other vulnerabilities. CVE-2018-17202 had the most difference between security group and generic group, which means the subjects in intervention group (security group) are more accurate than the generic group on detecting the specific vulnerability. The rest of the vulnerabilities almost had identical results, and we cannot say there is significant difference between two groups. Therefore, we can conclude that only the intervention in vulnerability CVE-2018-17202 succeed.

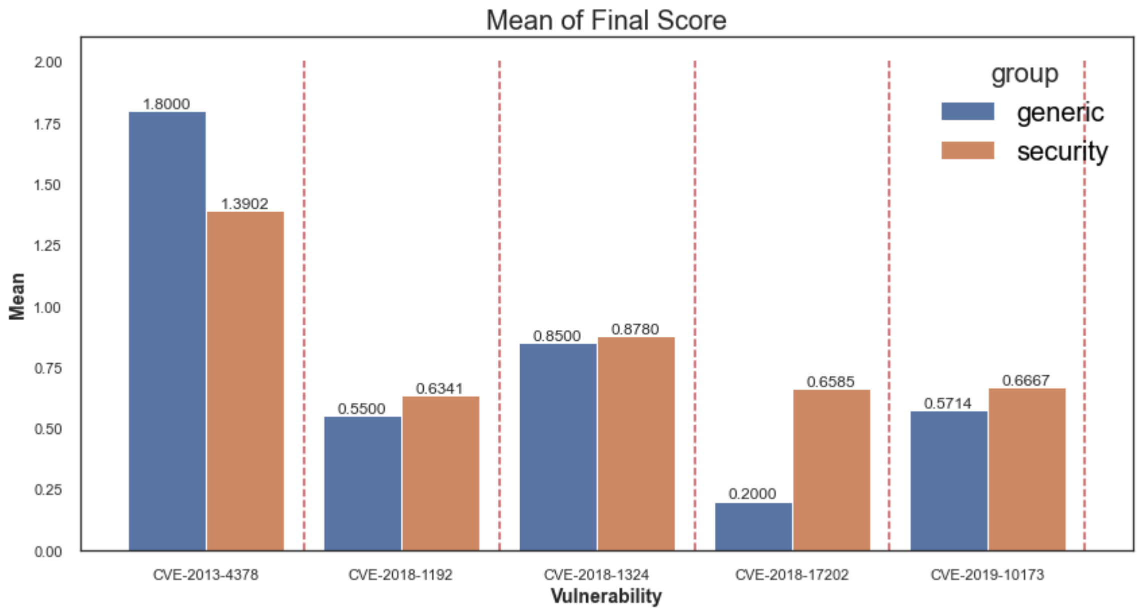

FIGURE 9

In our experimental setting, the influential rate represents the discrepancy between each subject's final score and the original score. A positive influential rate indicates that more subjects altered their answers to a more accurate one, while a negative influential rate suggests that more subjects modified their answers to a more erroneous one, indicating the APR tools succeed in misleading subjects' judgement. As the degree of influential rate increases, so does the efficacy of the APR tools. However, the lower the influential rate, the worse the actual effectiveness of the APR tools is.

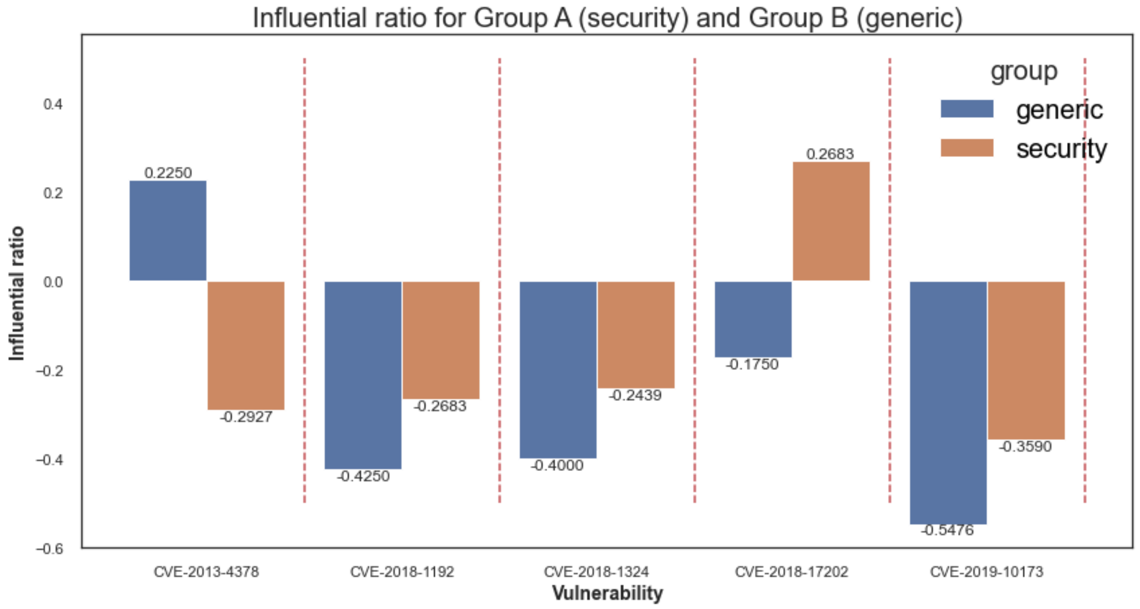

FIGURE 10

From FIGURE 10, we can observe that a lot of subjects are misled by the APR tools, no matter it is the actual security tool or not. However, the influential rate under each vulnerability is not very high. Note that with regard to the fact that only subjects with changed actions are taken into consideration in the influential rate, we can only analyze the success of intervention on this part of the subjects. If a subject received the suggestions from the APR tools but he still insisted on his original answer, then the implementation of the invention did not have any substantial impact on this part of the subject. If we ignore the subjects who chose to insist, we can deduce that **although the effect of the APR tools is mostly negative, the intervention does exert a relatively positive influence on the experiment results.** This can be inferred from the data presented in FIGURE 10, which indicates that the mean influential rate in security group is higher than the mean influential rate in generic group in four out of five cases. We will expound upon the significance of this finding in the subsequent section by employing statistical tests.

RQ2: Perception usefulness of the APR tools

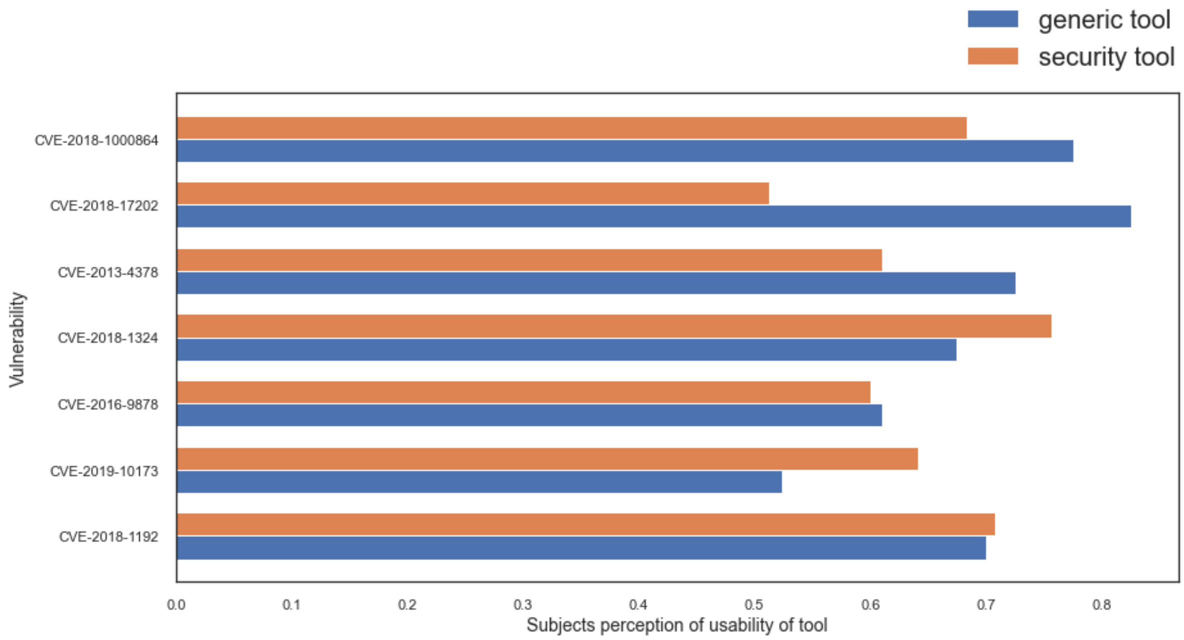

FIGURE 11

In FIGURE 11, we showed how subjects see usability of tools. The way we define the usability of tools is as follows. We checked how many subjects in each vulnerability insisted on their chosen patch after they were presented with a security tool patch, grouped by generic and security tools. Then we divided that number by all answers the subject gave on that group (bogus or security tool. Based on the data presented in FIGURE 11, it can be inferred that the perception of tool usefulness was reported by over 50% of subjects across all vulnerabilities, no matter which group they are in. Notably, in one instance, a tool was deemed useful by more than 80% of subjects. The average of perception is shown in the TABLE 7. As we can see, both type of tool have almost the same perception percentage, but generic tool has slightly higher average by 5% which means subjects tend to consider the tool useful when it is matched with their choice and does not really consider the fact that they might have made a mistake or the tool might be faulty or fake.

Tool type	Generic tool	Security tool
Perception Average	69%	64%

TABLE 7

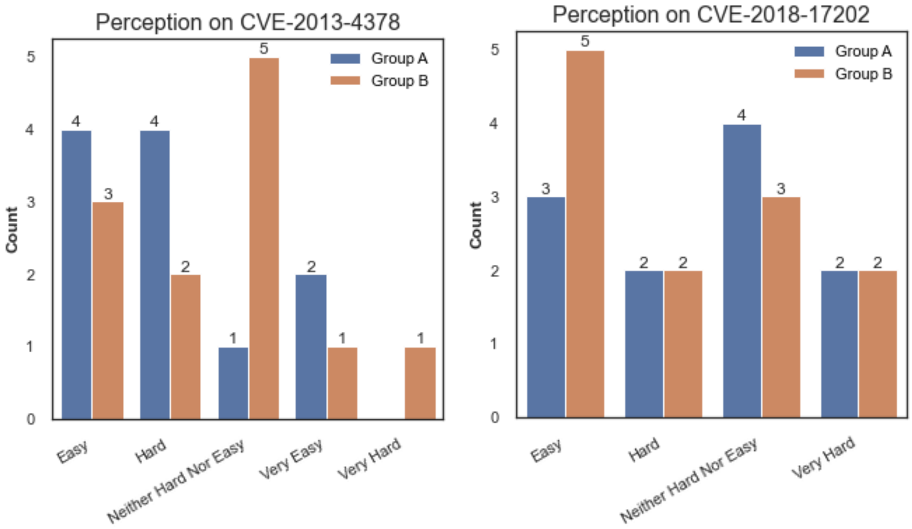

FIGURE 12

For the sake of comparison of the data between 2022 and 2023, we plot two perception bar charts using the data from last year. We select the most two remarkable vulnerabilities which is CVE-2013-4378 and CVE-2018-17202. We can find that the last year perception on the difficulty of these two vulnerabilities are positively correlated with the results in FIGURE 10.

9. Analysis Procedure and Results

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 statistical results.
Explicitly mention how you decided to analyze the data (a replication from past study, a new analysis observing different measures or using different statistical tests).

In this section, we will use statistical tests to verify our findings from descriptive analysis. We can not do any tests for part 1 as there is no control groups and no comparison can make. Therefore, we only conduct the statistical test for part 2 and RQ2 to determine if the actual effectiveness and perceived usefulness of the actual security tool differs significantly from the actual effectiveness and perceived usefulness of the actual generic tools.
Firstly, we need to decide the outcome measure we want to compare between the two groups. We choose the final score and influential rate, as they are the most valuable metrics which can directly reflect the difference between two groups.

Before conducting any statistical tests, we need to we check the normality of the data. We plot the histogram in FIGURE 13. The upper left subplot is the histogram for all data points. The rest three are the histogram of influential rate, final score and original score for CVE-2018-1192, which is randomly selected from all five valid vulnerabilities. We can clearly see that the upper two subplots follow the normal distribution, while the lower two not precisely follow the normal distribution. For the sake of explicitness and simplicity of comparison, we assume all data follows the normal distribution and the two groups have equal variances.

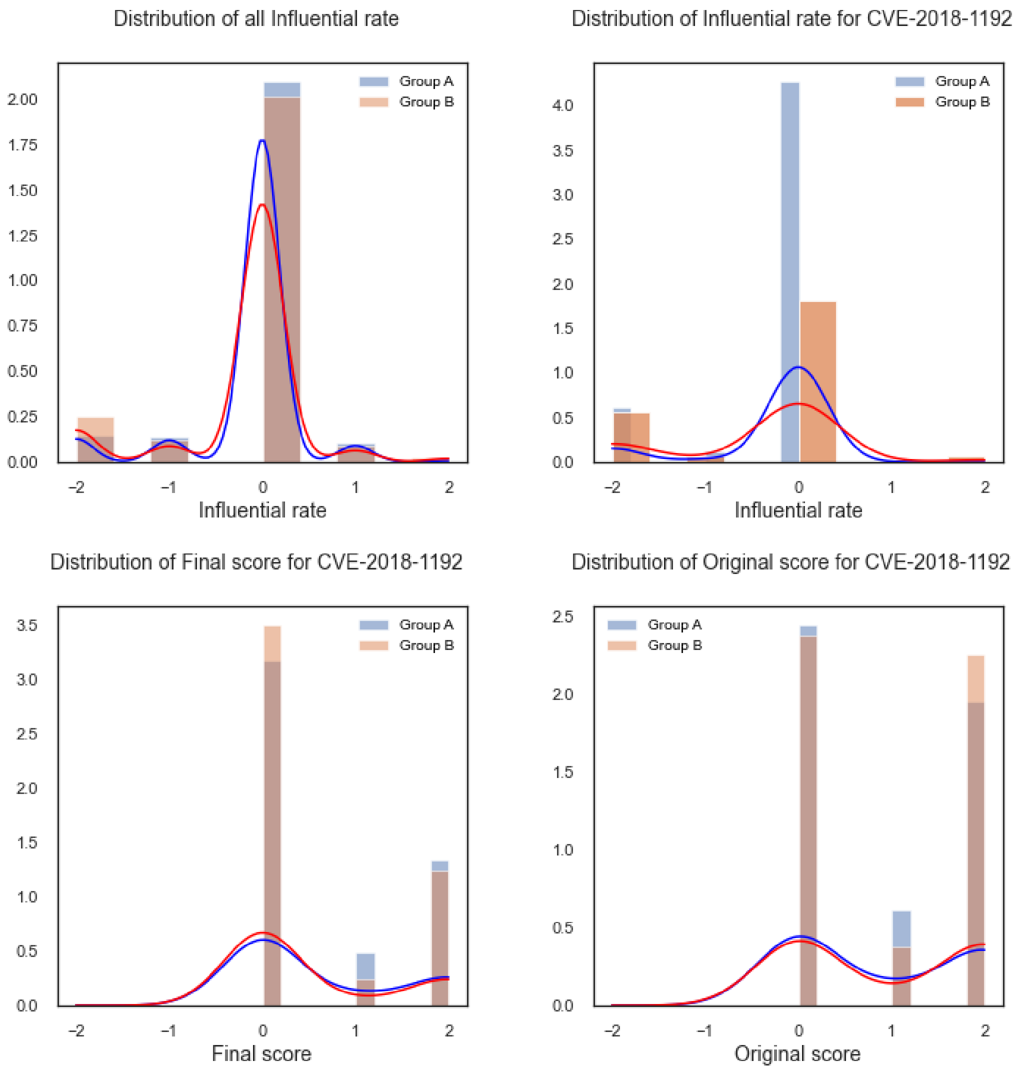

FIGURE 13

Now that we know final score and influential ratio are normally distributed, we can use t-test as a statistical test. The t-test, as a typical statistical, 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.

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** (influential rate / final score / perception). 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.

	Final Score		Influential Rate	
Vulnerability	t-value	p-value	t-value	p-value
CVE-2013-4378	4.077089	0.000108	4.954500	0.000004
CVE-2018-1192	-0.429438	0.668773	-0.870598	0.386611
CVE-2018-1324	-0.125820	0.900194	-0.761782	0.448459
CVE-2018-17202	-4.640163	0.000014	-4.768372	0.000008
CVE-2019-10173	-0.487577	0.627200	-1.051912	0.296047

TABLE 8

TABLE 8 shows the p-value and t-value of t-test for each group of security tool and generic tool per vulnerability. Taking the influential rate on CVE-2018-17202 as an example, the obtained p-value is 0.000008, which is greatly smaller than the significance level of 0.05. This means that we succeed to reject the null hypothesis and conclude that there is a significant difference between the means of the two groups at the 5% significance level. The t-value is -4.768372, which measures the size of the difference between the means of the two groups relative to the variable (influential rate) within each group. The negative sign indicates that the mean of Group B (generic tool) is lower than the mean of Group A (security tool). Meanwhile, the absolute value of the t-value is large enough to reject the null hypothesis. From the TABLE 8, we found that the two metrics final score and influential rate are correlated. Therefore, elaborating only one is enough. There are totally two p-values are less than the significance level of 0.05, respectively CVE-2013-4378 and CVE-2018-17202, indicating sufficient evidence to reject the null hypothesis. Besides, only 1 t-value is positive, indicating that the mean of Group B is lower than that of Group A for the other four vulnerabilities.

	t-value	p-value
Final Score	-0.578928	0.562868
Influential Rate	-1.187182	0.235655

TABLE 9

If we consider a subject's answer on one specific vulnerability as a data point and abandon or omit the information of who produce it, we then can merge all data points together only by the group (not also by the vulnerability) and conduct the t-test on overall data points. TABLE 9 shows that p-value in both two metrics are not small enough to reject the null hypothesis.

	t-value	p-value
Perception	0.951029	0.360352

TABLE 10

TABLE 10 presents the t-test result for the subjects' perception by using the data shown in FIGURE 11. Note that we also assume the data follow the normal distribution. The p-value in this metric not small enough to reject the null hypothesis, thus, we cannot say the perception of two groups differs significantly.

The results of these three tests are sufficiently to verify our aforementioned quantitative and qualitative analysis in Section 8. We conclude that **the actual security tool does help in identifying the vulnerabilities comparing with the generic security tool only in vulnerability CVE-2013-4378 and CVE-2018-17202, and in general, the difference is not significant.** We refrained from merging the data from last year and this year for the statistical test due to the variations in experimental setup procedures and insufficient important relevant information.

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.

The experiment had two primary goals. The first was to evaluate the effectiveness of Automated Program Repairs (APR) tools in helping developers identify and fix vulnerabilities, while the second was to determine whether human reviewers' decisions on the effectiveness of a security patch were influenced by the knowledge that it was generated by a specialized security tool. The report introduced a metric of success based on these goals and described the experiment's process, decision criteria, and analysis of data. A t-test was conducted on influential rate, final score, and subjects' perception to further analyze the data. The main findings elaborated in Sections 8 and 9 are summarized below.

1. APR tools effectively helped developers identify and fix vulnerabilities, with a mean accuracy of over 50% across all subject, vulnerability, and general levels.
2. Most subjects showed a high degree of consistency in their responses, indicating strong adherence to their initial selection.
3. The use of a specialized security tool only influenced the decision of less than 40% of human reviewers evaluating the effectiveness of the patch. For those influenced by the tool, it was found to be helpful in identifying vulnerabilities in specific cases, but the difference was not significant overall.
4. Subjects tended to consider the tool useful when it matched their choice, regardless of potential mistakes or faults in the tool. The difference between the perception of subjects who are suggested by generic tools or by security tool is not significant.

The study's limitations included sample size and subject background, with some limitations in the experiment design, such as all the patches produced by APR tools being incorrect for two vulnerabilities and only one vulnerability correctly identified by the actual security tool. The intervention implementation was critical to obtaining valid conclusions, and the determination of criteria for defining correctness was essential. We found that the randomization is done pretty good in this experiment when analyzing the Goal 2. There may be some possible errors regarding the correctness determination step for part 2 if our initial assumption is not correct. Finally, due to the absence of important relevant information, the report did not analyze or compare last year's data extensively.