Checking Compliance with Security Standard

Experiment Reproducibility Report - E-4

Security Experiments and Measurements

Vrije Universiteit 2023

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We hereby confirm that this assignment is our own work, is not copied from any other person's work (published or unpublished) and was not shared with any other person that could copy its contents.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 the experiment E4 is to explore the effectiveness of the automated rules (in the DeltaAICert tool) in facilitating the security compliance analysis process for analysts whether the automated rules. The experimental task involved performing the analysis once using the tool's automated rules (Group B) and once without them (Group A). The intervention is **the automated rules generated by the DeltaAICert tool**. Group B is the intervention group and Group A is the control group. Our study primarily focuses on measuring the impact of using these automated rules on reducing the manual effort required to achieve the same level of correctness in delta certification.

The DeltAICert tool serves multiple purposes, including aiding security evaluators in evidence gathering, facilitating delta evaluations, supporting evidence filtering, and enabling compliance with various security certification schemes. The tool's automation capabilities assist in the assessment process and generate reports and certifications. It is non-trivial to assess whether the implementation of automated rules in the DeltaAICert tool leads to a reduction in manual effort and results in efficient and effective security compliance analysis.

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.

RQ1: *Time* required to achieve *correct* delta analysis.

The correctness of the assessment consists of two parts.

1. The correctness of the final assessment for the requirement (AssessmentStatus).
2. The correctness of the files listed as related evidence to support the above assessment.

Note that the correctness of the assessment comment(s) may also be a possible aspect, but with regard to its subjective nature, we decide to not include this aspect in this report. The details of the criteria for these two parts are elaborated in Section 5.

Group-wise metrics:

1. The accuracy (correctness) among each group.
   (accuracy = correct_subject / all_subjects)
2. The mean of the accuracy calculated by different criteria for each group.
3. The mean, standard deviation, median of duration time among each group.
4. The mean, standard deviation, median of duration time only for the correct assessment among each group.

RQ2: The *perceived usefulness, usability, and effectiveness* of the DeltaAICert tool as collected by participants' questionnaires with Likert scale-like questions.

Specifically, in order to evaluate the perceived usefulness, usability, effectiveness and pace of the DeltaAICert tool, it is essential to assess how it helps developers in identifying certification compliance or compliance violations for new changes, assessing compliance, and speeding up the compliance analysis process.

To quantify the perception of the subjects, answers to the questionnaire are mapped to "1" to "5" accordingly. For example, there are questions about the usability of DeltaAICert, in which "Very Little" is mapped to "1" and "Extremely" is mapped to "5". Then we can apply statistical tests based on the quantified perception data, not only in a qualitative way. We may also be able to derive the conclusion whether our qualitative findings are statistically significant or not.

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 74 subjects attending the experiment while 36 of them are in group A and 38 are in group B. In this experiment, we are not provided with the past data, therefore, the past study could not impact our results.

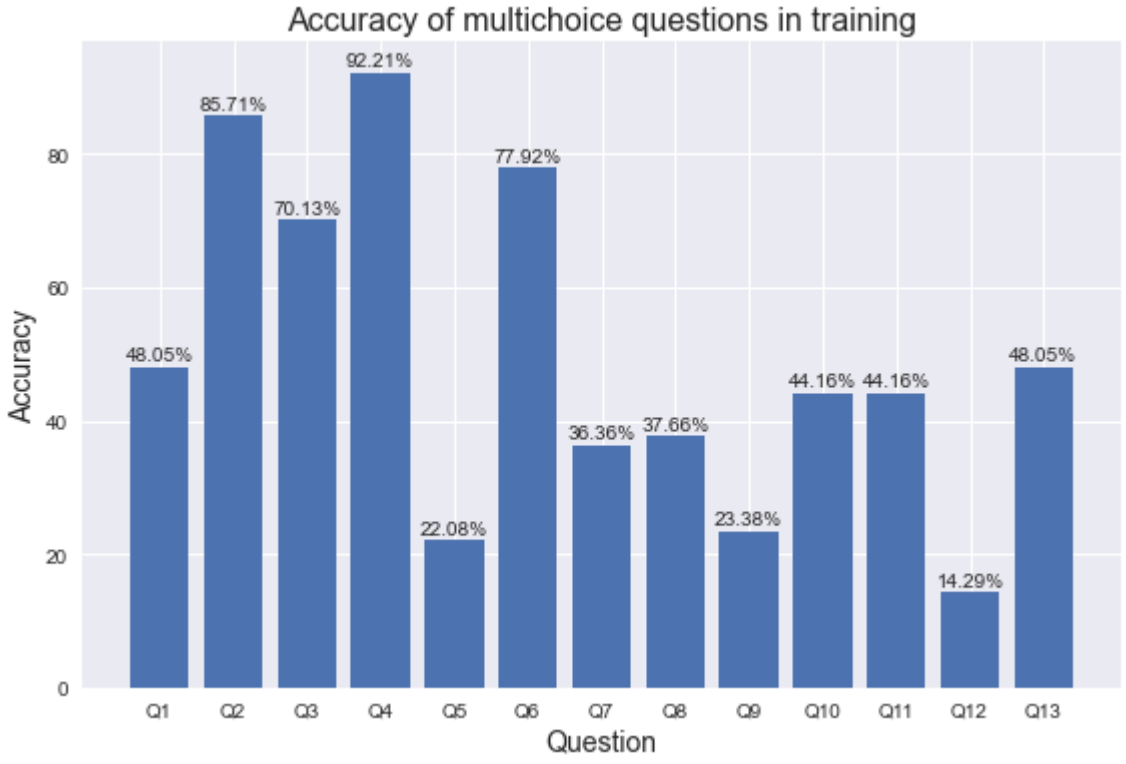

FIGURE 1 Accuracy of multichoice questions in training

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 were randomly assigned into two different groups and were provided with training videos and slides on security compliance analysis, mainly about how to use the DeltaAlCert. The training involved two groups that received different materials, where the materials for Group B included the introduction and instruction on the filter rules and assessment rules functions, while these functions were absent in the materials provided to Group A. After the tutorials, there were four questions to investigate the participants' background, as well as fourteen questions in the questionnaire to evaluate how the participants learned about the Piggymetrics and DeltaAlCert from the tutorial. Notably, the teacher maintains control over both groups by ensuring that each group was given the same amount of reading time, and both groups were asked the same questions in the end. The correctness of all participants' answers are shown in the FIGURE 1.

2. PARTICIPATED in an experiment

The second step is conducting intervention. After reading and watching all the material, the experiment is conducted. Although both groups were requested to evaluate the identical project, Piggymetrics project, each group received different task instructions. Participants in Group A were requested to evaluate the AUTH.6, while participants in Group B were requested to evaluate DATA.2. The details of the difference between two instructions are shown in TABLE 1. Two groups were presented with an updated version of the project and had to re-assess whether a particular security requirement is satisfied, considering the initial evaluation of the project. Specifically, the task of Group A involves observing evidence files, comparing changed files to their previous version to assess, and detecting a server-side throttling mechanism in the Piggymetrics program using the DeltAlCert tool. The task of Group B involves identifying keywords for logging function, observing associated evidence files, filter rule and its output, and making assessment to meet the requirement of not logging sensitive data in the PiggyMetrics program. Finally, all the participants need to select project files as evidence to support the assessment, close the evaluation, and generate the JSON and PDF reports.

	Group A	Group B
Requirements	AUTH.6	DATA.2
Requirement description	The remote endpoint implements a mechanism to protect against the submission of credentials an excessive number of times.	Security sensitive data (e.g., keys and certificates) and privacy sensitive data (e.g., PII) should not be written to log files.
Assessment results	Rate limiting in source code missing and the requirement failed.	Email addresses were logged; therefore, personal data was logged and the requirement failed.
Action	Assign the evidence files which contain relevant changes to the AUTH.6 requirement.	Run the preloaded filter rule by pressing the green arrow. Observe the output and the difference with the previous version of the output.

TABLE 1 The details of the difference between two instructions

3. ANSWERED question in the Qualtrics survey

After uploading the JSON and PDF into the Qualtrics survey, participants are asked to justify their final assessment answers and describe why they evaluated the delta certification as passed/failed for the requirement. Eventually, all participants are requested to fill out a survey, which was about evaluation of the whole task as well as some perceptual questions like their confidence in compliance assessment, perceived difficulties, and perceived the usefulness and usability of DeltaAICert tool. This evaluation serves as a critical indicator of the participants' comprehension of the material and their intuitive understanding of the experiment tasks and their results, making it a crucial aspect of the experiment.

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. As the past data is not provided, therefore, we can compare our data with the past study. In order to determine the positives and negatives based on the ground truth, we can use two types of validation procedures: 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 matched the JSON and PDF reports generated by participants with the ground truth report, and identified our own criteria to obtain positives and negatives (more details will be elaborated in Section 6).

6. CONDUCTED the statistical analysis

We proposed various statistics measures for evaluation, specifically various statistics measures for the duration time of the correctness results, plotted the data for easier comparison and elaborated the results quantitatively and qualitatively. After that, we analyzed the perception data of all subjects and conducted quantitative analysis.

7. CONCLUDED the results and experiment

We conducted the analysis procedure by using a statistical test, and compared the results with the finding of statistical analysis. Note that, 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.

Some limitations of this experiment design include the subjects' background and the randomization. The experiment does not account for the confounding variables that may affect the ability of the subjects to execute the task, 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 randomly divided into two groups.

One of the key limitations is the difficulty on finding a valid success of measures. In this experiment, two groups were provided with two different materials and executed two different tasks. Although the intervention is successfully implemented as subjects in Group B can attain the automated rules generated by the DeltaAlCert tool, while the subjects in Group A cannot, the two objectives for two tasks are totally different. One focuses on *A throttling mechanism should be implemented on server side* and another focuses on *No sensitive data is written to logs*. Since this experiment did not control for all variables, we cannot determine the success of the experiment by comparing some common effective metrics, such as accuracy. The time required to achieve correct delta analysis is really an indirect metric, which contains many are many uncontrollable confounding variables. These variables could include differences in the skills and experience of the individuals conducting the experiment, and variations in the performance and availability of the computer tools involved. For example, there were some participants who had issues with the tool, and it took a long time for them to fix the issue. All the fixing time was included in their experiment time. Besides, from the FIGURE 1, we can observe that the accuracy rate of more than half of the questions was less than 50%, which shows that the performance of the students' training is very poor. Many students don't really understand how to use the DeltaAlCert tool to check compliance. Therefore, the problem of ineffective training procedure can seriously affect the time they spend on tasks. In addition, we only consider the time taken by subjects with correct results, while subjects with incorrect results are excluded as outliers. This operation may lead to too much valid data being regarded as outliers, resulting in too few experimental samples, and thus affecting the validity and statistical power of the experimental results.

Therefore, the time required to achieve correct delta analysis contains a lot of noise, and it is important to use multiple metrics to evaluate the success of the experiment. However, the perception of a subject is a very subjective metric and may be affected by their unique cognitive and learning styles, and since the experiment did not control for these factors, it could limit the reliability and validity of the results. Thus, we decide to take the accuracy among each group as a reference to ensure a comprehensive and accurate assessment of the effectiveness of the security measures being tested.

Another critical point is the possibility of confounding variables (e.g. the effectiveness of DeltaAlCert tools), which may impact the results of the experiment. For instance, the effectiveness of DeltaAlCert tools may depend on the type of tasks, the participants' level of expertise on JAVA, etc. With regard to different tasks for two groups, the effectiveness may diverse. Therefore, these factors were not controlled for in the study.

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.

In our experiment, the "artefacts" referred to subjects' JSON and PDF reports generated by the DeltaAICert tool. After reading the training material and task instructions of two groups and randomly comparing some PDF files with two ground truth files, we determine the ground truths of two groups are as follows.

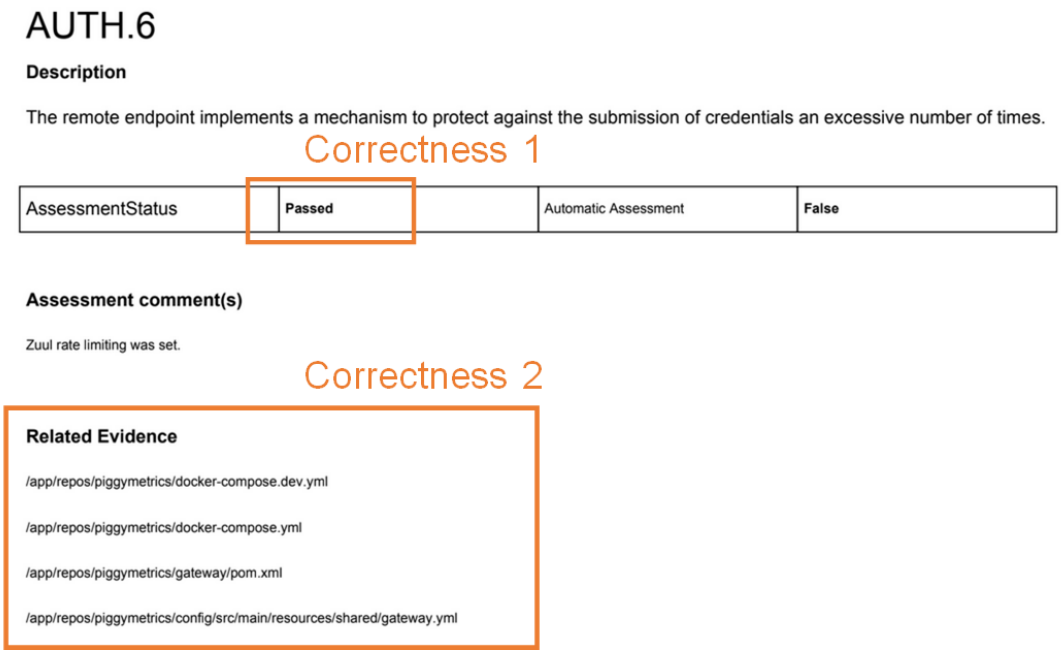

FIGURE 2 Group A Ground Truth

FIGURE 3 Group B Ground Truth

FIGURE 2 and FIGURE 3 separately show the ground truth of Group A and Group B. Correctness 1 represents the first part of correctness of the assessment we mentioned in Section 2 (possible answers for subjects are "Passed", "Failed", "Inconclusive"). While

Correctness 2 represents the second part of it. To unambiguously define positive results, a subject's AssessmentStatus artefact must precisely match Correctness 1; otherwise, the result is negative or incorrect. As for the Related Evidence, various criteria emerge.

1. Exact Match: When a subject's Related Evidence artefacts exactly matches the ground truth of Related Evidence (Correctness 2), the subject's artefacts is defined as correct. Everything else is considered as incorrect.
2. All Match: When a subject's Related Evidence artefacts contains all the ground truth of Related Evidence (Correctness 2), the subject's artefacts is defined as correct; otherwise, it will be considered incorrect.
3. Half Match: When a subject's Related Evidence artefacts contains half of the ground truth of Related Evidence (Correctness 2), the subject's artefacts is defined as correct; otherwise, it will be considered incorrect.

The restriction ranges from strong to weak. The first one may be overly restrictive, and the last one might be too optimistic. Too strict may result in less noise but smaller sample size, on the contrary, too weak may result in larger sample size but more noise. Note that, both rules of Correctness 1 and Correctness 2 must be satisfied to finalize whether the subject's report is correct or not. Only satisfying one of the rules cannot say the report is correct. After we determine the correctness, we can use the duration time of each subject with correct report to calculate the mean, geometric mean and standard deviation of duration time among the group as the measures of success. We will compare the measurement results of all three criteria in Section 8.

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 positives and negatives. With regard to the large quantity of artefacts (74 PDF files or JSON files), we decided to apply an automatic procedure, which is a python script, to evaluate subjects 'artefacts. The automatic procedure contains three steps, separately artefacts preprocessing, criteria selection and correctness determination. Notably, as the JSON files are the structured version of PDF files, we decide to use the JSON files in our ground truth generator.

Artefacts preprocessing: In the `"RawData"` folder, we have two sub-folders, separately `"Q34"` and `"Q35"`. Normally, `"Q34"` should contain all the reports with JSON format and `"Q35"` should contain all the reports with PDF format. However, as two of the subjects does not precisely follow the instruction, we primarily need to read all the files in two folders and redistribute them according to their format. Since the files in the folder are not classified according to their group, we also need to manually divide all JSON files and PDF files into two categories. Note that, the filename of the report submitted by each subject corresponds to his ID (`ResponseId`), we then associate each subject's group according to their ID.

Criteria selection: After the preprocessing, we first extract the ground truth from two JSON files and transferred the ground truth into the wanted data type (dict). Subsequently, we extract the corresponding data points from all subjects' artefacts into the wanted data type (dict). We then map the ground truth dictionary with the subject dictionary to obtain four types of result dictionaries `"only_assessment_match"`, `"only_evidence_match"`, `"both_match"` and `"no_match"`, we also dub them as group criteria in Section 8. The `"only_assessment_match"` dictionary contains the data of the subjects whose artefacts only match Correctness 1 and do not match Correctness 2. The `"only_evidence_match"` dictionary contains the data of the subjects whose artefacts only match Correctness 2. The `"both_match"` dictionary contains the data of the subjects whose artefacts both match Correctness 1 and Correctness 2, which is the intersection of `"assessment_match"` and `"evidence_match"`. The `"no_match"` dictionary is the complement of the union of the above three dictionaries. Each dictionary contains the information of group ID, mapped subject ID (`ResponseId`) and their duration time (`duration_time`). We implement all three criteria separately in this step. At last, we have 3 big dictionaries separated by criteria, each of which contains 4 sub-dictionaries.

Correctness determination: We determine that only the data points in the `"both_match"` dictionary are the correct artefacts (the reason is elaborated in the precious section), the rest of three dictionaries are used for comparison. Therefore, after the aforementioned steps, we have obtained the correct artefacts under the three criteria respectively. We then use the `duration_time` of each data point in `"both_match"` dictionary to count the metric of time. Finally, we use the outputs to calculate various statistics measures of the metric. Besides, we calculate the proportions of subjects of `"both_match"`, `"no_match"`, `"only_assessment_match"`, `"only_evidence_match"` under the three criteria defined in the previous section, and compared them to analyze which criteria is more suitable for statistic tests.

There may be some possible errors regarding the correctness determination step, as some subjects may select a lot of related evidences (much more than the number of the correct). If their results match all the correct related evidences, we still determine their artefacts as correct, which is not very accurate. Additionally, the determination of the ground truth may cause some errors. This is because the ground truth provided by the teacher is four files (two of them are JSON files and two of them are PDF files), after we carefully read the training material and task instructions, we randomly sampled some subjects' artefacts in PDF format and manually compared them with the ground truth in PDF format. Then, we found the difference between the artefacts and ground truth, and we determined that only the AUTH.6 is the ground truth of Group A and DATA.2 is the ground truth of Group B. We chose the method of manual comparison to find the ground truth because each JSON file is very large and contains numerous keys. It is very inefficient to compare by writing code, and may easily produce wrong results. But applying manual comparison may also result in missing some other possible ground truth.

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 adopt a manual validation procedure for our algorithm. The procedure is similar to the one we used to determine the ground truth. Firstly, we randomly sample 5 PDF format artefacts from Group A and 5 PDF format artefacts from Group B in a random sequence. The reason why we select the PDF format is that the data in JSON files are all structured data, which are more easily to read by the computer, therefore, we use the JSON files in the automatic procedure. In the manual validation procedure, all the artefacts will be evaluated by human assessors. Therefore, using the PDF format artefacts is clearer and more friendly to the human assessors. Then, 10 artefacts are assigned to each group members, which are two independent human assessors. All the artefacts are unmarked at the start. The preliminary evaluation will 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”), “0” refers to “incorrect” while “1” refers to “correct. In this step, we go through all the content in the PDF file and compare it one by one with the content of the ground truth PDF file to see whether they were exactly the same. If the same, we grade it “1”. If not, then, my teammate and I work together to come up with common scoring guidelines, where both Correctness 1 and Correctness 2 must be satisfied. Note that, we decide to apply the second criteria for Correctness 2 in this procedure. Once the common scoring guidelines are established, we two continue scoring the remaining artefacts according to these guidelines. If the scores grade by two human assessors are the same, then we judge whether it is a good sample or a bad sample based on its score. If the conflict emerges, we will choose to re-score it until the score is identical.

After scoring all the sampled artefacts, we obtain the good samples (correct) and bad samples (incorrect). We compare our results with the results generated by the automatic procedure, if matched, then it is a truly good or truly bad sample. We find that all the artefacts are either truly good samples or truly bad samples, which means the correctness assignment of our aforementioned automatic analysis is all correct. The results are shown in TABLE 2. Therefore, the error rate is 0. With regard to the low error rate, we decide to not apply any error correction operations for the results generated by our script.

	Group A	Group B
Truly Good	3	4
Truly Bad	2	1
Error rate	0	0

TABLE 2 Results of manual validation

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 preprocess the data and filter out the invalid data, e.g. replicate data. Then, we visualize the duration time to find the outliers. Luckily, according to FIGURE 3, all subjects in two years finished in the requested time. We can observe that subjects in Group A generally spent more time on the experiment than subjects in Group B regardless of median, mean and interquartile range. We can refer that the automatic rules help subjects to check the compliance from a holistic perspective.

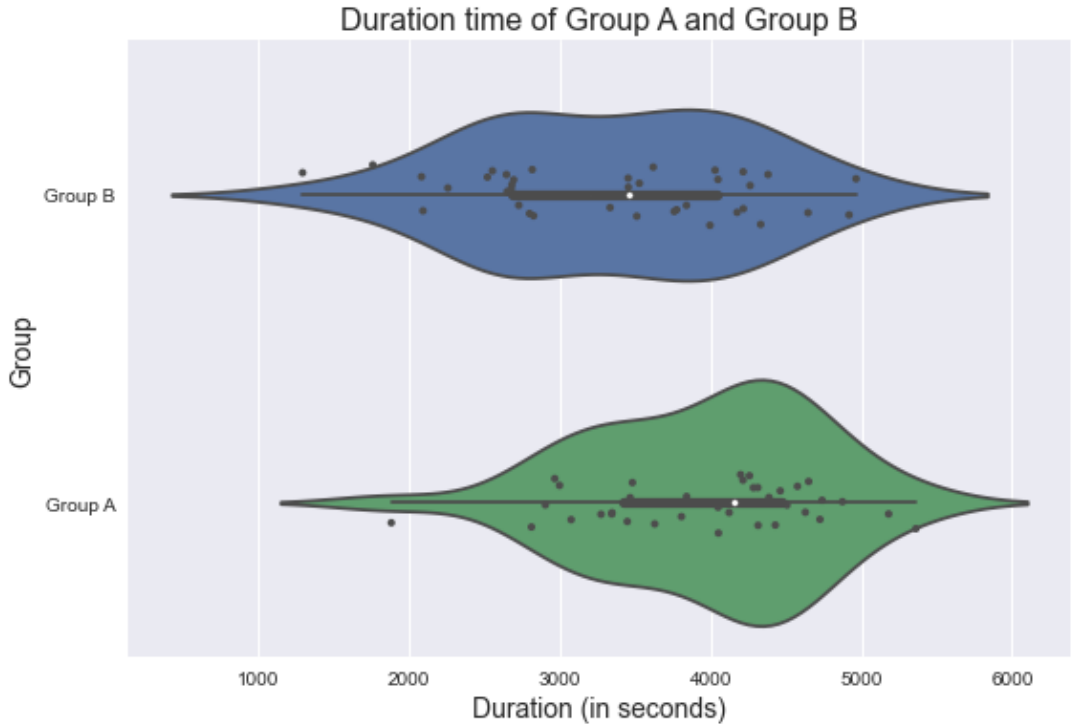

FIGURE 4 Duration time of all subjects

Duration time	Group A	Group B
Mean	3962.416667	3336.921053
Standard deviation	753.594183	898.868910

TABLE 3 Mean and Std of Duration time

Following the preprocessing stage, we establish three specific criteria and subsequently implement an automated procedure to ascertain the positives and negatives for each subject. These values are then utilized for further analysis of the data, using basic descriptive statistical measures, accuracy. After that, we extract the duration time spent by the positive subjects and conducted descriptive analysis of it in detail.

In this section, we conduct interpretation and analysis based on two research questions. When analyzing RQ1, we conduct in-depth quantitative and qualitative analysis of the duration time in subject-level and group-level based on different criteria. When analyzing RQ2, we quantify the perceptions and confidences of all subjects and analyze the data

quantitatively. Note that we didn't merge the two years data together because the last year data is not given to us.

RQ 1: Group-wise Metrics

After conducting the ground truth generator, we automatically separate all the data into four dictionaries, respectively `"only_assessment_match"`, `"only_evidence_match"`, `"both_match"`, `"no_match"`. The explanations on what each dictionary represents are elaborated in Section 6. In the two figures below, we compare the counts of subjects and the overall accuracy among each group by applying three criteria.

	both_match	only_assessment_match	only_evidence_match	no_match
Criteria 1 Exact Match				
Group A	10	16	4	6
Group B	21	1	13	3
Criteria 2 All Match				
Group A	12	14	7	3
Group B	21	1	16	0
Criteria 3 Half Match				
Group A	20	6	7	3
Group B	21	1	16	0

TABLE 4 The count of subjects in four dictionaries under three criteria

Accuracy	Criteria 1	Criteria 2	Criteria 3
Group A	0.277778	0.333333	0.555556
Group B	0.552632	0.552632	0.552632

TABLE 5 The overall accuracy of two groups under three criteria

We can easily observe that the restrictions of three criteria range from strong to weak. When applying the Criteria 1, very few subjects in Group A match both the Correctness 1 (assessment) and Correctness 2 (evidence), as well as a lot of subjects only match the Correctness 1. When the restriction becomes weaker, the situation changes as more subjects in Group A attain both matches. Interesting, when observing the values of Group B, we don't find the identical rule. No matter what the restrictions are enhanced or weakened, the same amount of subjects attain both matches and only assessment matches.

Another rule that is different between the two groups is that Group B is characterized by a higher number of subjects who achieve only evidence match rather than only assessment match, while as for Group A, it is the opposite. However, it should be noted that matching

assessment is generally considered to be an easier task than matching evidence. This is because assessment questions typically have three choices, and only one of them is correct, while evidence questions are multiple-choice questions with an uncertain number of correct answers. Specifically, for the questions in Group B, there are a total of eight correct answers, and if we apply Criteria 1, all eight answers must be exactly matched, which means that the subject's answer must be eight and all correct. Intuitively, the question of evidence is expected to be more challenging than the question of assessment. The results of Group A align with this intuition, whereas the results of Group B do not. We can infer from the above findings that the intervention (automatic rules) has a great effect on helping subjects to find the related evidence accurately, as most subjects in Group B successfully and exactly found all the evidence.

Because only the data points `"both_match"` dictionary is what we will use in the further analysis, therefore, we calculate the overall accuracy among each group which is the number of data points in `"both_match"` divided by the number of all data points. As shown in TABLE 5, when comparing the overall accuracy of three criteria, we find that the values of accuracy of Group B are exactly the same no matter which criteria is applied, on the contrary, the values of accuracy of Group A become larger as the restriction becomes weaker. Even in Criteria 3, the accuracy of Group A is slightly higher than the accuracy of Group B. This finding can verify that the intervention has a great effect on helping subjects to check compliance. For the sake of clarification, we will mainly compare the results of Criteria 3 in the following analysis, as the accuracies of two groups under Criteria 3 are quite similar to each other.

In addition to the three correctness criteria defined in Section 5, there are also four group criteria, which are correspondence to the four aforementioned dictionaries, separately `"only_assessment_match"`, `"only_evidence_match"`, `"both_match"` and `"no_match"`. Note that these four criteria won't affect the definition of correctness, and only subjects in the `"both_match"` group are defined as correct when conducting the analysis on time.

As we can see from the FIGURE 5, in Group A, if we consider the subjects attain `"only_assessment_match"`, the overall duration of subjects in this group criteria, have the lowest duration, which means they did the task correctly faster compared to the `"only_evidence_match"` and `"both_match"` groups. For the `"only_evidence_match"` group, the mean is getting higher. This finding is consistent with the intuition that it takes more time for subjects to read the evidence and determine the correct evidence compared to the assessment task. Moreover, it is interesting to note that in Group A, the subjects who correctly identified the evidence and assessed compliance took the longest time on average. However, the average duration of time taken by subjects in the `"only_evidence_match"` group is almost the same as that of the `"both_match"` subjects in Group A.

In Group B, the interesting thing is that only one subject only detected the assessment correctly, the rest were either only detected evidence correctly or both detected evidence and assessment correctly. Also, same as Group A, the `"only_evidence_match"` group took more time to finish the task. But unlike group A, in the "both_match" group, the median

and mean of duration time is less than the "only_evidence_match" group. The experimental analysis reveals that Group B exhibits consistently lower task durations across all three conditions compared to Group A. This indicates that the intervention group, which utilized the automated rules, generated by the DeltaAICert tool, were able to complete the task in less time.

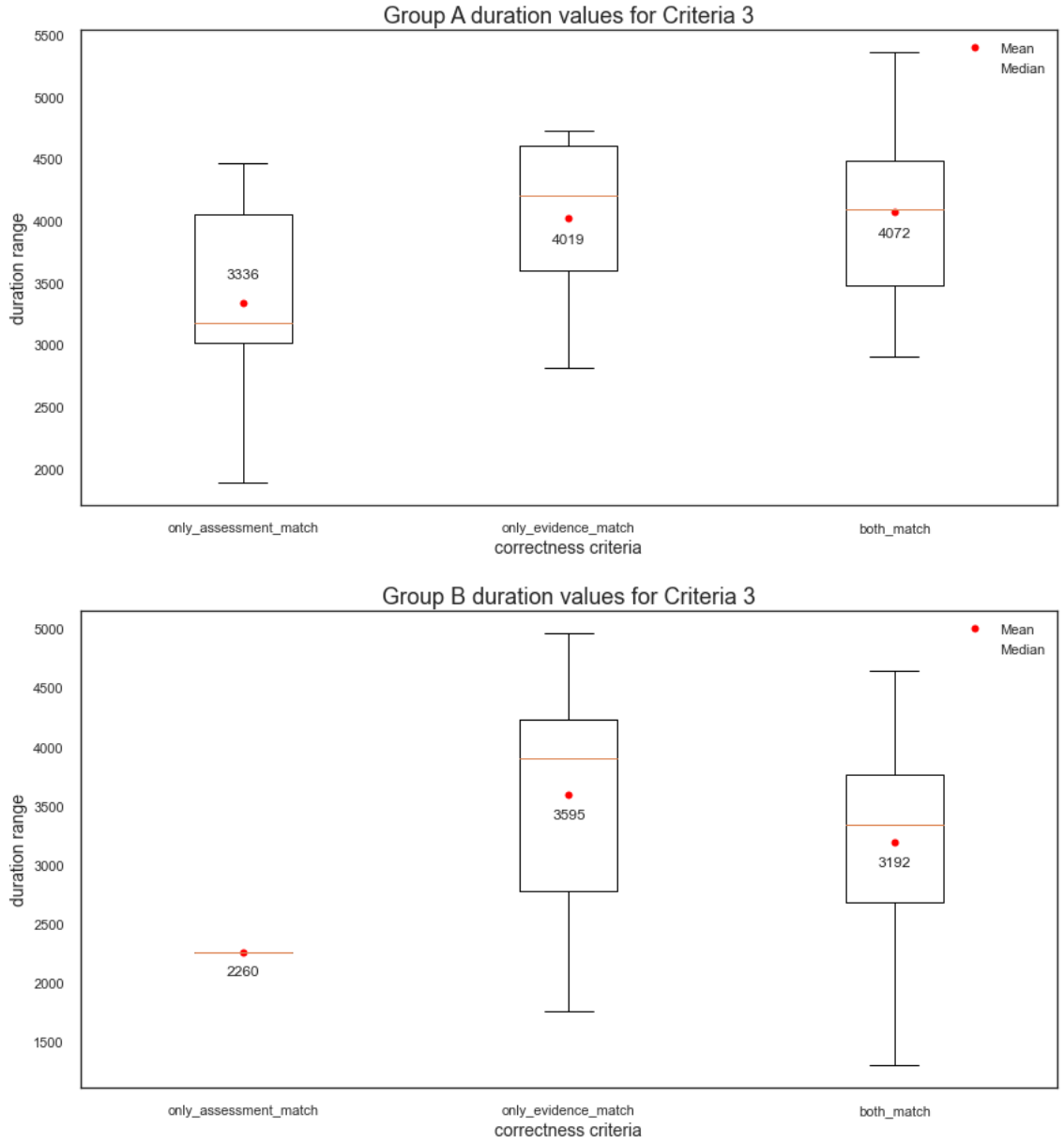

FIGURE 5 Duration time values under Criteria 3

In summary, our aforementioned analysis implied that automated generated rules actually help subjects to speed up the process. This finding also aligns with the perception analysis of pace. When analyzing the perception of pace, it is observed that the number of subjects who reported perceiving a significant improvement in pace ("A lot") was higher than those who reported perceiving moderate or minimal improvement. This finding supports the notion that the DeltaAICert tool was perceived by subjects as helpful in speeding up the process of analysis. Overall, these results highlight the potential benefits of incorporating automated rule generation tools in experimental analysis and demonstrate the positive impact that such tools can have on both objective task metrics, accuracy and duration time.

RQ 1: Subject-wise Metrics

FIGURE 6 shows the duration time values of Group A under three Criteria in the subject-level. We only plotted the values of Group A as the values of Group B are exactly the same under three different criteria. We can observe that, along with the restriction of criteria weakening, the median and interquartile range of duration values become smaller. Likewise, we can roughly infer that the time subjects spent is directly proportional to their accuracy.

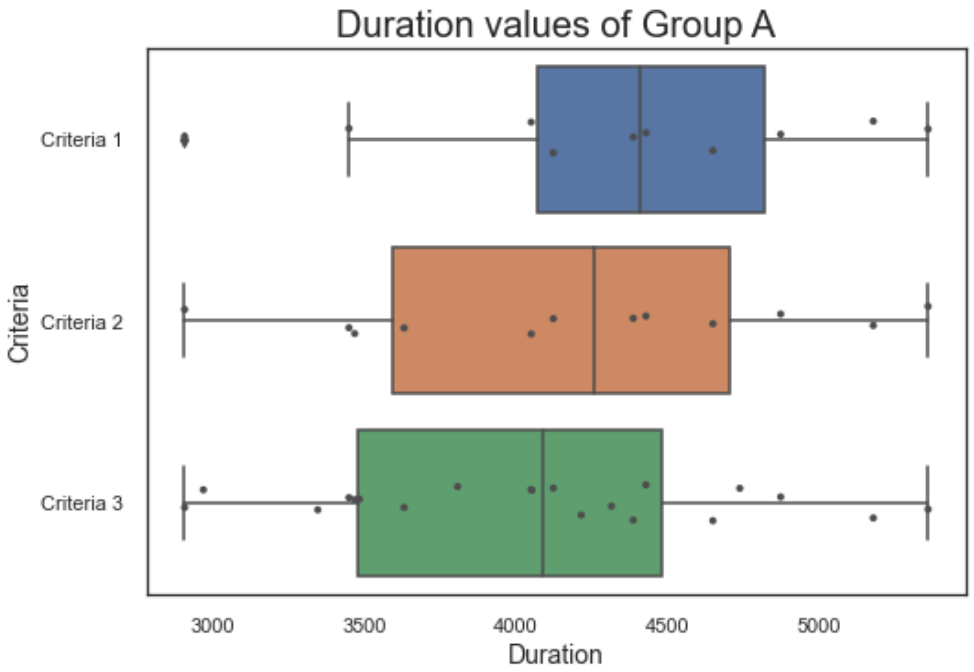

FIGURE 6 Duration time of Group A until correctness under three Criteria

After conducting a comparison between the group-wise metrics using three criteria, we chose to only apply Criteria 3 for comparing the duration values in subject-level.

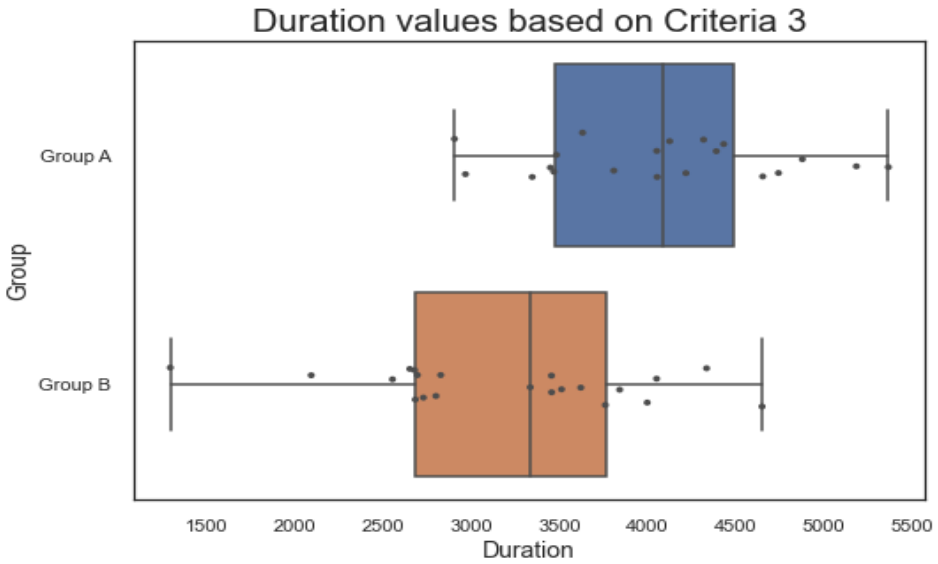

FIGURE 7 Duration time until correctness under Criteria 3

FIGURE 7 only plots the duration time until correctness of the delta analysis under Criteria 3. We can conclude that subjects in Group B completed the task more quickly while still maintaining a high level of accuracy than the subjects in Group A regardless of the median, minimum, maximum and interquartile range.

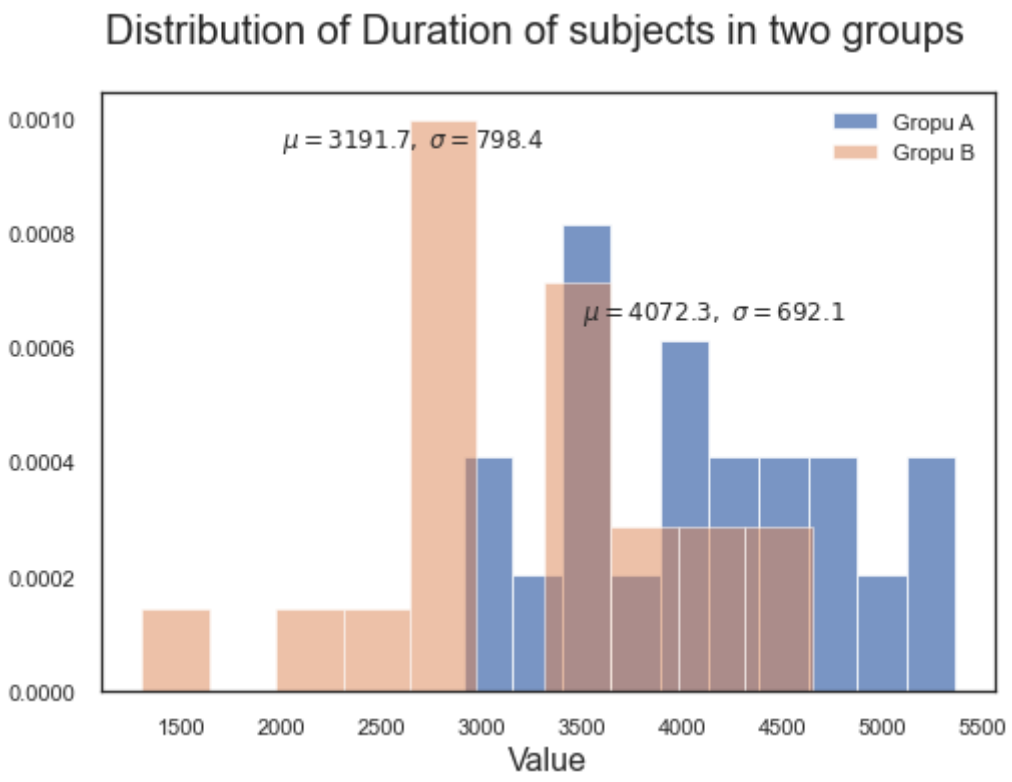

FIGURE 8 Distribution of Duration time values under Criteria 3

FIGURE 8 is the histogram on the duration time until correctness of the delta analysis of two groups under Criteria 3. It reflects the overall distribution of the experimental results. We notice that the data points of Group A roughly follow the normal distribution, but the data points of Group B seem not precisely following the normal distribution.

RQ 2: Perception

In this part, we firstly analyzed four different factors including usability, usefulness, effectiveness and pace of the subjects' perceptions. Usability is related to the question "**How did you find the usability of the DeltaAICert tool?**". Usefulness is related to the question "**To what extent did you find the DeltAICert tool useful to identify certification compliance or compliance violations for the new changes in the repository?**". Effectiveness is related to the question "**To what extent did the DeltaAICert tool help you in assessing compliance?**". Pace is related to the question "**To what extent do you agree that the DeltaAICert tool helped you in speeding up the compliance analysis?**". All questions have five possible answers, separately "**Extremely**", "**A Lot**", "**Neither Little Nor A Lot**", "**Little**" and "**Very Little**". Based on the degree of perception, we assigned "1" to "5" to the above choices accordingly.

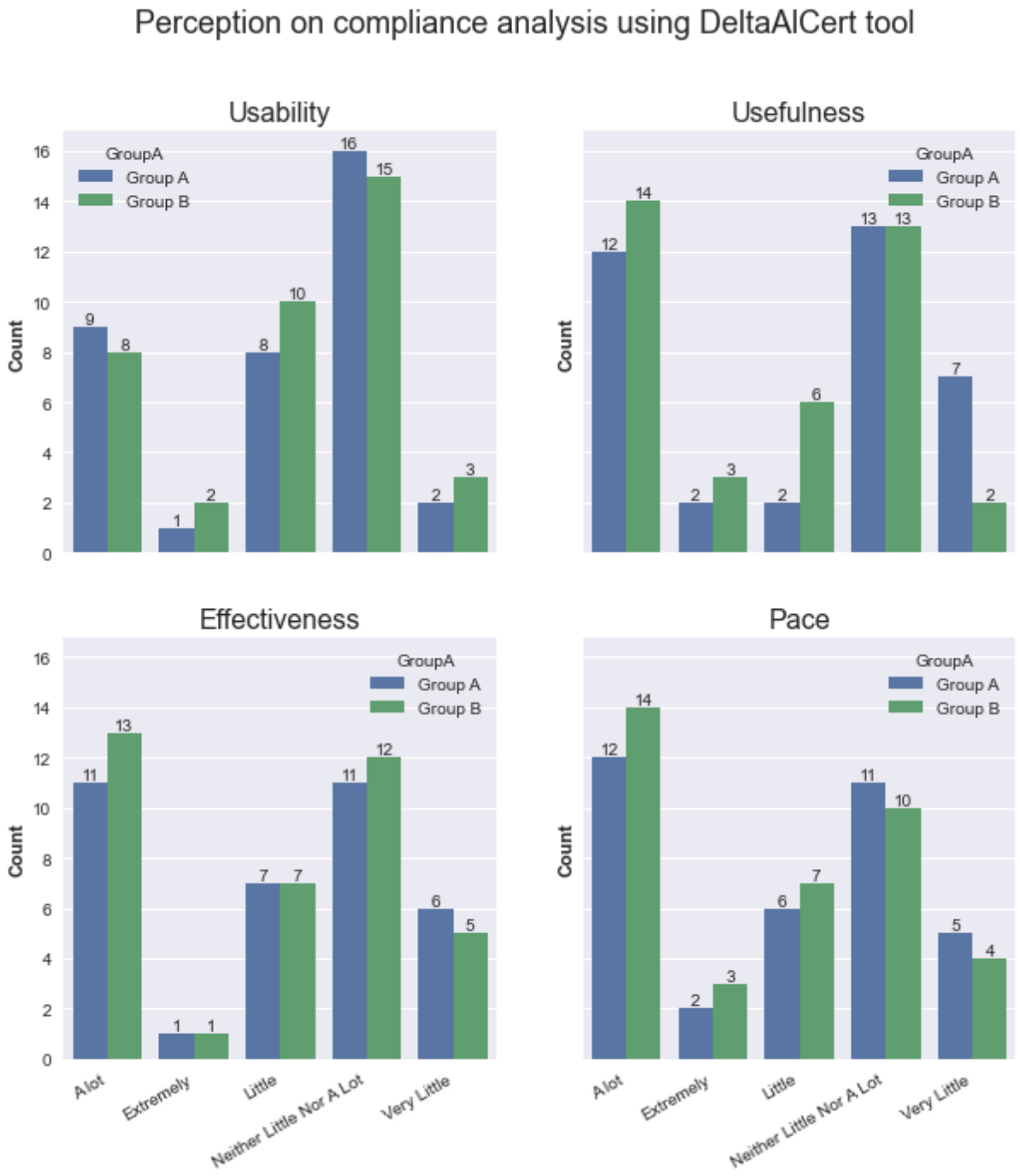

FIGURE 9 Perception of Usability, Usefulness, Effectiveness and Pace

As we can see from FIGURE 9, in all four factors, majority of subjects had a perception of “Neither little Not A lot” and “A lot”. This suggests that subjects generally held a positive perception of the DeltaAICert tool. In terms of Usability, most subjects neither found the tool usable nor unusable. But In Usefulness and Effectiveness, the number of subjects in “Neither Little Nor Alot” and subjects in “A lot” were almost the same. This suggests that subjects perceived the tool to be both useful and effective in assisting with compliance analysis. Additionally, in regards to the pace factor, the group that perceived a significant improvement in pace ("A lot") had a higher number of subjects. This implies that subjects believed that the DeltaAICert tool significantly helped to speed up the process of compliance analysis.

Upon comparing the two groups, no significant differences were observed. While a slightly higher number of subjects in Group A perceived the tool as usable compared to Group B, no substantial differences were found in terms of effectiveness between the two groups. However, there were some notable differences between the groups in terms of usefulness and pace. Nonetheless, the difference in perceived effectiveness between the two groups was insignificant. We will verify this finding in the next section by applying the statistical test.

Overall, these results suggest that the DeltaAICert tool was generally well-received by the subjects and perceived as useful, effective, and capable of improving the pace of compliance analysis.

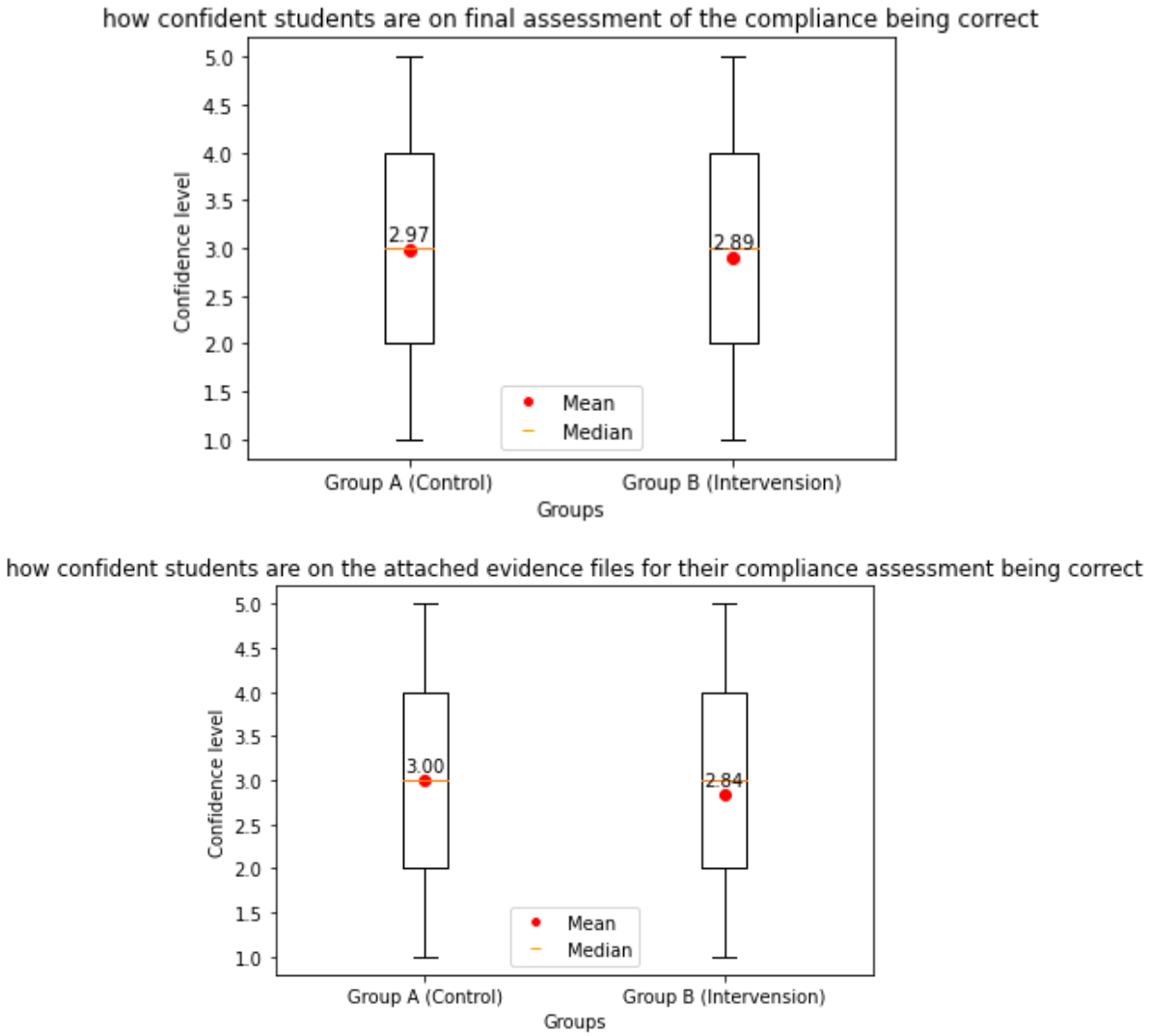

FIGURE 10 Confidence on the final assessment

FIGURE 10 is two box charts, which depicts the confidence on the final compliance assessment of two groups of subjects. It is apparent from the charts that both groups exhibited almost equal levels of confidence in their answers, with not much difference between the two groups. However, the box chart for Group A shows slightly higher levels of confidence in their assessments compared to Group B.

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.

In this section, we will use statistical tests to verify our findings from descriptive analysis. Firstly, we need to decide the outcome measure we want to compare between the two groups. We choose the duration time and subjects' perceptions, as they are the most valuable metrics which can directly reflect the difference between two groups.

RQ 1

As we cannot distinguish whether the metric of time follows the normal distribution or not from FIGURE 8, we choose to perform the Shapiro-Wilk test to double-check the normality of the data.

For doing statistical analysis, we applied Shapiro-Wilk test on duration time of the experiment for control group (Group A) and intervention group (Group B) separately to see if data is normally distributed. The null hypothesis of the Shapiro-Wilk test is that **the specific dataset is normally distributed**. The results are shown in the following table:

	Group A	Group B
p-value	0.883961	0.607142

TABLE 6 The p-value of the Shapiro-Wilk test for RQ1

As we can see, p-value for both group is higher than threshold that usually is considered in academic (0.05) which indicates that we failed to reject the hypothesis. Then it is proven that data is normally distributed for both groups.

After we proved the data is normally distributed in both groups, we can perform the t-test to see how different the two groups of data are on the metric of time. The t-test calculates a t-value and a p-value, which is compared to a critical value or a significant level to determine if the differences in the means are significant.

Before the analysis, the null hypothesis needs to be assumed, stating that there is **no significant difference between the two groups in the metric of time**. The significance level, alpha, is 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.

	t-value	p-value
Criteria 1	3.812502	0.000664
Criteria 2	3.599479	0.001096
Criteria 3	3.765400	0.000548

TABLE 7 The results of t-test for RQ1

TABLE 7 shows the p-value and t-value of t-test for each group under three different criteria. Taking the result of Criteria 3 as an example, the obtained p-value is 0.000548, which is greatly smaller than the significance level of 0.05. This means that we succeed in rejecting 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 3.765400, which measures the size of the difference between the means of the two groups relative to the variable (duration time) within each group. The positive result indicates that the mean of Group A is greater than the mean of Group B, which is consistent with the results shown in FIGURE 7. Meanwhile, the absolute value of the t-value is large enough to reject the null hypothesis. From TABLE 6, we find that all three p-values are less than the significance level of 0.05, indicating sufficient evidence to reject the null hypothesis.

Based on these results, we can conclude that the difference between Group A and Group B on duration time until correctness of the delta analysis is significant. This conclusion verifies that **the intervention does help subjects to speed up the process.**

RQ 2

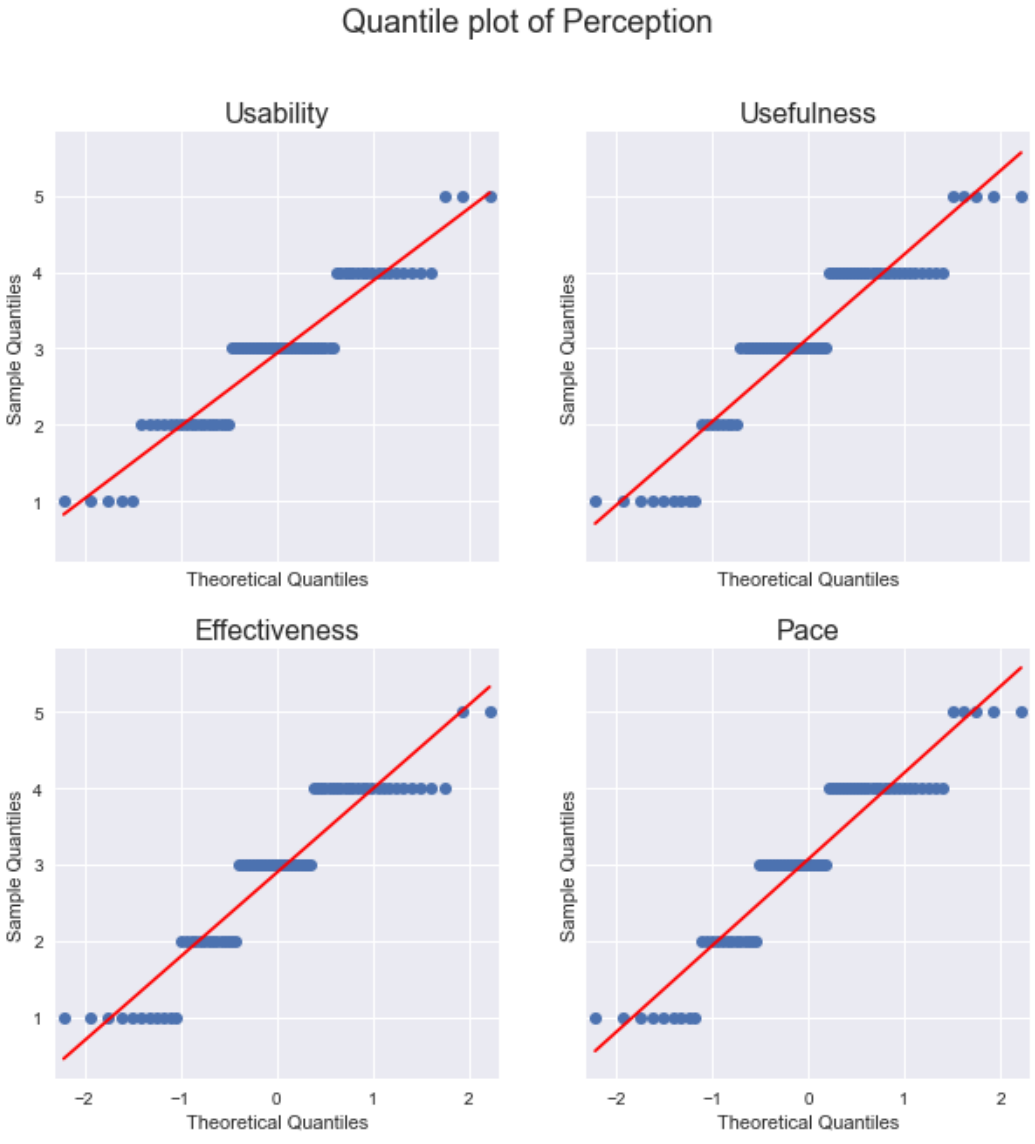

FIGURE 11 Quantile plot of Perception

QQ plots are an essential tool in experimental analysis, as they allow us to visually compare two probability distributions by plotting their quantiles against each other. Specifically, QQ plots provide a way to determine whether two sets of data are approximately normally distributed or not. If the two distributions being compared are identical, the points on the QQ plot will fall exactly on the line y=x, indicating that the data points follow a normal distribution. However, if the distributions are not identical, the points on the QQ plot will deviate from the line y=x, indicating a departure from normality.

p-value	Group A	Group B
Usability	0.003035	0.006768
Usefulness	0.000276	0.002802
Effectiveness	0.001781	0.000836
Pace	0.002657	0.002271

TABLE 8 The p-value of the Shapiro-Wilk test for RQ2

In our case, we can observe from FIGURE 11 that data samples are too sparse. Therefore, QQ plots may not accurately assess the normality of the data. The results in TABLE 8 also verify that the perception data aren't normally distributed. In this case, alternative methods, such as non-parametric tests, may need to be considered to analyze the data. Therefore, in order to accurately analyze the data and determine if there is a significant difference in perception between the two groups being compared, despite the lack of normal distribution in the data, we decide to conduct Mann-Whitney test. Instead, it tests for a difference in the location of the two groups by comparing the ranks of the data values between the two groups.

The null hypothesis of Mann-Whitney test is assumed as **automatic rules are not perceived as useful, or no significant difference between the two groups in the four metrics of perception.** The significance level, alpha, is also set at 0.05.

Perception	p-value
Usability	0.342832
Usefulness	0.786276
Effectiveness	0.669421
Pace	0.691464

TABLE 9 The p-value of the Mann-Whitney test for RQ2

Similar to the analysis of t-test, as shown in TABLE 9, all p-values are larger than 0.05, which indicates that the evidence is insufficient to reject the null hypothesis. Thus, we cannot say the perception of two groups differs significantly.

The results of these two tests are sufficient to verify our aforementioned quantitative and qualitative analysis in Section 8. We conclude that **automated generated rules do help subjects to speed up the process and the influence is significant, while in terms of their perception of the usability, usefulness, effectiveness and pace of DeltaAICert tool on checking compliance, the difference is not significant between two groups.**

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.

This experiment aims to explore the effectiveness of the automated rules in the DeltaAICert tool in reducing manual effort required to achieve the same level of correctness in delta certification, with Group B being the intervention group and Group A as the control group. The study focuses on measuring the impact of using the automated rules on facilitating the speed of security compliance analysis process. The study introduced several metrics of success based on these goals, separately time until correctness of delta analysis and perception of usability, usefulness, effectiveness and pace of the tool. After describing the experiment's process, decision criteria, and descriptive analysis of data, two statistical tests, respectively t-test for RQ1 and Mann-Whitney test for RQ2, were conducted to further analyze the data. The main findings elaborated in Sections 8 and 9 are summarized below.

1. The intervention of automated rules in Group B had a significant effect on helping subjects accurately find related evidence.
2. In terms of duration time, Group B consistently exhibited lower task durations across all three criteria compared to Group A.
3. Subjects in Group B completed the task more quickly while still maintaining a high level of accuracy than the subjects in Group A.
4. The DeltaAICert tool was generally well-received by the subjects and perceived as useful and effective on compliance analysis.
5. There is a statistically significant difference in the duration time between positive subjects in the two groups.
6. There is no significant difference between the two groups regarding their confidence in the final compliance assessment and their perceptions of usability, usefulness, effectiveness and pace.

In general, automated generated rules significantly speed up the compliance checking process, but there is no significant difference between the two groups in their perception of the DeltaAICert tool's usability, usefulness, effectiveness, and pace.

While this experiment provides some insight into the effectiveness of DeltaAICert in improving the speed of compliance checking, there are several limitations that need to be considered. The lack of control over confounding variables such as participants' expertise in the field and level of expertise in JAVA, as well as the subjective nature of perception metrics, may impact the reliability and validity of the results. Additionally, the use of time as

the primary metric may not be sufficient to fully evaluate the success of the experiment. Future studies should consider controlling for these factors and using multiple metrics to ensure a comprehensive and accurate assessment of the effectiveness of the security measures being tested. Finally, due to the absence of past data, this report did not analyze or compare last year's data extensively.