Uppsala University

Department of Informatics & Media

Usability Test Report 1st Draft

*Usability Test Plan*

*Frimanson Per, Jonathan Joan, Salah Omar-Alfred, Sameer Mishra & Torres Padilla Juan Pablo*



Course: Human-Computer Interaction: Usability Evaluation Methods, 7.5c

Date: 20181123

Contact: Per.Frimanson.7925@student.uu.se,

Joan.Jonathan.6216@student.uu.se,

Omar.Salah.4511@student.uu.se,

[Sameer.Mishra.8872@student.uu.se](mailto:Sameer.Mishra.8872@student.uu.se) &

[Juanpablo.Torrespadilla.6297@student.uu.se](mailto:Juanpablo.Torrespadilla.6297@student.uu.se)

# 1 Introduction

The objective for this usability test was to determine if there is a significant difference in the time it takes to complete the same management tasks in both Discord and Slack, to determine which application has as a more effective interface and how a user’s experience with the software affects the time needed to complete tasks within it. For this purpose, quantitative data was gathered through a controlled laboratory experiment in which the participants’ screens were recorded while performing the same tasks in both applications. These recordings were then coded in order to determine the start and finish time for each task, which was then used to calculate the actual task completion time.

Statistical tests were used to analyse the data, in particular, independent t-tests were performed using the experience of the participants in each application as a grouping factor. From these results, some conclusions were hinted at but present low statistical significance.

# 2 Data collection methods

The study required test subjects with previous experience with common communication software. For this purpose, students were fitting since most students regularly use computers. The sampling method used was convenient sampling, since the time frame required readily available and cost effective samples.

The data was gathered from 9 users, two of which had previous experience with both Slack and Discord, one who had experience with neither, three that had experience with Slack but not with Discord, and three who had experience with Discord but not with slack. The user was categorized as experienced if they used the software at least weekly.

No users were categorized as outliers and all of the users data was used in the analysis, although some users did not complete all tasks. Results for tasks that were uncompleted were not included in the data analysis, this was done automatically since SPSS only calculates the results on applicable data points.

# 3 Data collection and description

Task completion time was collected by first coding each video in order to determine the starting and finishing time for each task. Using this information, the time spent on each task was calculated, measured in seconds. In total, nine people performed the test, yielding sixteen values of task completion time, eight for each application.

The amount of wrong clicks for each task were also recorded in each video. Although this is not central in trying to prove or disprove the null hypothesis, the data can be somewhat of an indicator of the complexity of a task or of how easy it is to navigate the application’s interface.

# 4 Data analysis methods

The resulting data was analysed using independent t-tests, grouping the TCT values of each task by the subject’s experience with either of the applications, which was a yes or no value, the experience value was coded as 1 for ‘yes’ and 0 for ‘no’ for the SPSS analysis. This initially yielded two independent t-tests, one for each application. This allowed for a comparison on how experience with each software can affect the subject’s ability to efficiently/effectively complete certain tasks. Additionally, the mean task completion time of all tasks for each application was calculated. Using this data, another two independent t-test were carried out to determine the difference in general task completion time across the applications, again grouped by the experience each subject had with each application.

The intent was to also include an analysis of the mean task completion time for both subjects that had experience in both applications and that had no experience in either. The objective of this was first to account for different levels of experience with the two applications by having different combinations of experience. And to determine the possible impact on task completion time. However, due to the low size of the sample this was not possible to perform, as there were not enough data points showing said characteristics.

# 5 Data Analysis

For all t-tests in the following section S refers to Slack and D refers to Discord. For example S1\_TCT, represents the mean task completion time for task 1 in Slack. Additionally, S\_EXP and D\_EXP refer to the groupings for Slack and Discord experience respectively.

## T-test 1 grouping by Slack experience (1,0):

When looking at T-tests performed with Slack Experience (tables 1 and 2) as the independent grouping variable (which can contain the values Yes or No, indicating experience or lack thereof), there is a trend and slight indication that having experience in Slack can reduce the time taken to complete tasks in both Slack and Discord. This can be noted in the reduced mean TCT’s as observed for certain Slack tasks such as 4,5,7 and 8 This indication was even noted in the time taken to complete Discord tasks in 1,3 and 5. Despite this however, the lack of experience in Slack indicates that it did not necessarily hinder the time taken to complete certain tasks as observed in the mean times taken to complete Slack tasks 1,2,6 and Discord tasks 2,4 and 7. However, some results, such as those for Discord tasks 4 and 7 presented increased TCT for users with Slack experience. This can be potentially explained due to the very different process needed to complete that specific task across the applications.

The independent samples test performed for T-test 1 computed the mean differences in the time taken to complete tasks by the independent grouping variable Slack Experience with Levene's test for equality of variances computed as well. The outcome of Levene’s test indicates that equal variances are to be assumed when considering the results. Moving onwards, the tabulated results indicate that none of the differences in the mean taken to perform the tasks were significant across the different level of Slack experiences. This can be seen in the ‘Sig 2 tailed’ column of the table where none of the 2 tailed p-values are below 0.05 (which indicates the statistical significance of the difference in means). This means that within the small sample size, the difference in the means occurred by chance.

## T-test 2 grouping by Discord experience(1,0):

With regards to the second T-test performed with Discord Experience (tables 3 and 4) as the independent grouping variable. The same indications were noticed that having experience in Discord helped for some tasks but not necessarily for other tasks. For example, having Discord experience (1) helped with Slack tasks 1,2,3 and 6 and Discord tasks 2,3,4,6,7 but not significantly in the other tasks. An independent samples test was also performed for T-test 2 which computed the mean differences in the time taken to complete tasks by the independent grouping variable Discord Experience with equal variances assumed from Levene’s test. Likewise, the results also indicate that none of the differences in the mean taken to perform the tasks were significant across the different level of Discord experiences since none of the 2 tailed p-values were below 0.05.

## T-test 3 mean TCT’s for Slack tasks and Discord tasks grouped by Slack experience(1,0):

The results, as seen on tables 7 and 8, show a reduction on the mean task completion time for all the Slack tasks of about 13.4 seconds and an increase of 2.38 for Discord for users with previous experience in Slack. This shows that experience in Slack doesn’t translate into lower task completion in Discord, which as discussed earlier can be potentially explained by the several differences between the interfaces of both applications. However, the 2 tailed significance value is over 0.4 for Slack tasks and 0.8 for Discord tasks, which marks its statistical insignificance.

## T-test 4 mean TCT’s for Slack tasks and Discord tasks grouped by Discord experience(1,0):

In this T-test (tables 5 and 6), a reduction in the mean completion time for Slack tasks was not observed for experienced Discord users, so having experience in Discord did not necessarily translate to a faster task completion time for the Slack tasks. However, a marked improvement was observed in the mean it took to complete Discord tasks for experienced Discord users (24.4 seconds as opposed to 37.4 seconds). However, with the 2 tailed significance values being above 0.921 for the mean task time for Slack and 0.177 for the mean task time for Discord, the results therefore also become statistically insignificant.

# 6 Results

The goal of this study was to measure the difference it took to complete the same type of tasks across both Discord and Slack in order to compare the effectiveness of both of the applications. Therefore, to measure effectiveness across the applications, the following hypothesis was formulated:

There is no significant difference in the *time taken to complete the same tasks* across Discord and Slack (with varying levels of experiences in those applications).

The data gathered from the tests do indicate to a falsity to the null hypothesis and a trend behind the data point to a possible rejection of the null hypothesis, this can be seen in the Q-Q plots to some variables as some of them meet the prerequisite assumptions of a T-test, that is to say, that normality is observed in some of the gathered data of task completion time and somewhat close to 0.05 p-values for certain variables such as the D\_Mean. Ultimately, however, the null hypothesis at this point of time in the studycannot be rejected due to the high two-tailed significance values of the independent samples test, and the high p-values of the data gathered for each of the TCT variables for Slack and Discord. Therefore, a conclusion cannot be reached to if there is a significant difference in the time taken to complete the same tasks across both of the applications. This does not necessarily mean that the UTP test has been a pointless pursuit, but that following the recommendations proposed in later sections of this document may lead to a definitive answer to the null hypothesis.

# 7 Reflection

For the data to be representative of the targeted audience, online communication software users, the sample size would have to be bigger and without a normally distributed data set the prerequisites for our analysis methods were not met. A bigger sample size would also mean more diverse data. In the current set, while the users experience with the two programs was the same, there was a lack of users without experience in either, or with experience in both.

A bigger participant size would also allow for different samples for Discord and Slack, which would reduce the effect experience in one has on task performance in the other. By switching the order of the programs in the task list this was somewhat mitigated, but separate samples would reduce this factor further.

While convenience sampling suited this report because of tight timeframe and low resources, it introduces bias to the data. The sample may under-represent certain subgroups in the population. The groups relationships with some of the users could have effect user performance.

Some of the tasks the users performed were in hindsight considered redundant. The solutions to task three and four are located in the same place in both applications, meaning that when the user had completed one of them, they often solved the other right away. One of these task could have been removed because of this.

Task seven was meant to highlight that only one of the programs give a hint to how to notify a channel or user in a server. But it turned out that the users either already knew how to do this and completed both tasks very fast, or they didn't and took a long time to complete the task or didn't manage to complete the task at all. Since the tasks solution relied so heavily on whether the user knew about this command or not it didn't convey if the software provided usability, just that they had the same feature.

# 

# 8 Discussion

One of the main things learnt during the project was the importance of having a decent sample size and a well designed experiment. Additionally, it would be useful to include qualitative data, for example, with the use of post-test questionnaires or just gathering the participant’s impressions of the software after the test, since a purely quantitative approach used during this project didn’t seem to provide enough insight into any possible usability issues.

Furthermore, we have learned that having a proper usability test plan is important since it helps with consistency. Preparation is key since a lot of time can be saved by doing things correctly the first time. The pilot test was very useful since without a lot of mistakes would have been made, possibly rendering multiple user tests useless. Time management was also important, transcription took a long time and because of the timeframe such things needed to be considered.

Due to the reliance on statistical tests to reach any conclusion, the lack of participants was a detriment to the experiment. And due to the sampling method, it is hard to draw any generalizable conclusions since the sample was not rigorously selected. Thus, while some preliminary conclusions were drawn, their validity cannot be asserted. However, finding test participants was a challenge throughout the test and the reason why convenience sampling was used.

If the test had to be redone, the biggest change made would be to have fewer and better thought out tasks. As discussed earlier, a post-test questionnaire to gather more qualitative data would have allowed for a more detailed view of both applications. Finally, more effort would have been put towards having a bigger and more thought out sample.

# 

# Appendix A: Test plan

## Summary

The usability test will compare the communication apps Discord and Slack, both of which have similar features for managing communities and chat channels. The test will focus on the features that both applications share, such as creating and managing chat channels and rooms. As well as managing user roles or privileges, searching the channels, messaging specific groups. The test objectives are to measure and quantify how long it takes to complete these features/tasks and compare them between Discord and Slack using statistical analysis to measure the difference in effectiveness between these similar features of both Discord and Slack.

## Description of the system

Discord is an application originally made to help communicate via voice chat, however it has evolved to be used as a community hub. Slack on the other hand is targeted mostly to organizations, for employees to communicate between them.

Aside from minor differences, the interfaces of both applications are very similar. With a section of the screen devoted to showing the list of channels or servers, a big part of the screen devoted to the chat box and another part where other users in the server are displayed.

Discord has the additional option of having voice rooms, a feature that is not present in slack and thus will not be tested for.

## Target User Group

The target user group will be people within age 20-30 years that aren't regularly using either Slack or Discord before the test. However due to time constraints and difficulty of finding test participants, users with no experience in both software packages were accepted. To counterbalance the potential effects of learning how to do a task in one software before the other, the order in which users tested each application was switched for every user.

## 

## Data analysis methods

Since the data gathered will be quantitative (measuring the time taken to perform the tasks and the ordinal level of experience of the user using the application), descriptive statistics such as the mean, mode and standard deviation will be used using SPSS to analyse the data and draw conclusions from the data.

Since there is more than one dependent variable. Independent t-tests can be used to measure the difference in means between TCT among the different levels of experience for Slack and Discord.

TCT will be gathered and factored in a 2x2 matrix with the level of experience as shown in the table below:

|  |  |  |
| --- | --- | --- |
|  | Slack | Discord |
| Slack Experience | TCT/WC | TCT/WC |
| Discord Experience | TCT/WC | TCT/WC |
| Experience with both | TCT/WC | TCT/WC |
| No experience | TCT/WC | TCT/WC |

## 

## Test tasks

### Discord tasks

1. Create a new channel with a suiting name
2. Generate a invite link and paste it in the open text document
3. Changing the role of the user “Joan” to Administrator
4. Kick the user “Joan” from the server
5. Rename the channel you created to something else
6. Pin any message in any channel
7. Notify every user on the channel
8. Delete the channel you created

### Slack tasks

1. Create a new channel with a suiting name
2. Invite a new user, either by sending an invite to “[temporary email address]” or by generating and pasting a invite link in the open text document
3. Change the role of the “Omar” to Administrator
4. Remove the user “Omar” from the workspace
5. Rename the channel you created to something else
6. Pin any message in any channel
7. Notify every user on the channel
8. Delete the channel you created

# 

## Measures

In this study there are two independent variables and two dependent variables:

|  |  |
| --- | --- |
| **Independent variables** | The software used for the tests (Slack or Discord) |
| Users experience with similar systems |
| **Dependent variable** | Task completion time (TCT) |
| Number of wrong clicks (WC) |

The user experience variable will be gathered during a pre-test questionnaire, while the software used is determined during recruitment. During the test a TCT will be recorded per task completed.

## Test sessions

The tests will be conducted using a desktop computer where the test subject will have access to either Slack or Discord. Independent of the application to be used, the test subject will be given an account and a server where they can perform the tasks.

Before the test begins, the test subject will take a pre-test questionnaire to determine their level of experience with either Slack or Discord. If the test subject says they have experience with Discord then they will test Slack, and vice versa. During the test, one of the group members will have access to a user account on the application, to be able to respond to the channel invitations sent by the test subject.

Additionally, another person from the group will be the moderator, answering any questions the test subject may have and preparing the server and channels for the user to test.

The test sessions will be structured as follows:

### Participant preparation (5 minutes)

The participant will first review and sign a consent form to get permission to perform the test and use their data. They will then be introduced to the test, its purpose and how it will be conducted.

### Test (20 minutes)

First the user will fill the pre-test questionnaire from which it will be determined if they will use Slack or Discord for the rest of the test. The actual test then consists of eight tasks for each application, for a total of 16 tasks. Users are asked not to use a think aloud process, since this could alter the time needed for them to complete a task, rather users are asked to state when they are starting and finishing any given task, this is to have more precision when coding the resulting videos.

## Pre/post test questionnaires

The pre test questionnaire will contain two ordinal questions regarding the users’ level of experience with both Discord and Slack. The pre test questionnaire will determine whether users are experienced with either applications. It will also determine if the user has experience with using an online service to manage a community chat. If the users are somewhat, or very experienced with one of the applications, then the users will test the other application and vice versa, e.g: a somewhat/very experienced Discord user will test Slack and a somewhat/very experienced Slack user will test Discord. Users who are not experienced with either Discord or Slack will test either one of them based on the needs of the experiment, i.e: to appropriately match the populations for both samples.

## Coding Scheme

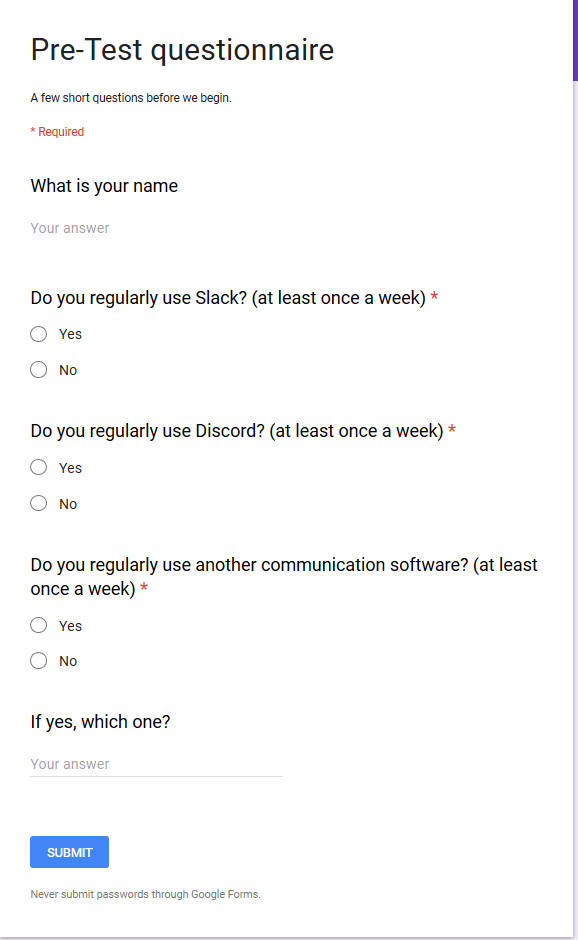
The BORIS application was used to code the recorded videos. The following coding scheme was used and applied to the videos:

|  |  |  |  |
| --- | --- | --- | --- |
| **Behaviour Type** | **Code** | **Description** | **Category** |
| State Event | ST1 | Slack Task 1 | Task |
| State Event | ST2 | Slack Task 2 | Task |
| State Event | ST3 | Slack Task 3 | Task |
| State Event | ST4 | Slack Task 4 | Task |
| State Event | ST5 | Slack Task 5 | Task |
| State Event | ST6 | Slack Task 6 | Task |
| State Event | ST7 | Slack Task 7 | Task |
| State Event | ST8 | Slack Task 8 | Task |
| State Event | DT1 | Discord Task 1 | Task |
| State Event | DT2 | Discord Task 2 | Task |
| State Event | DT3 | Discord Task 3 | Task |
| State Event | DT4 | Discord Task 4 | Task |
| State Event | DT5 | Discord Task 5 | Task |
| State Event | DT6 | Discord Task 6 | Task |
| State Event | DT7 | Discord Task 7 | Task |
| State Event | DT8 | Discord Task 8 | Task |
| Point Event | WC | Wrong Clicks | User Interaction |

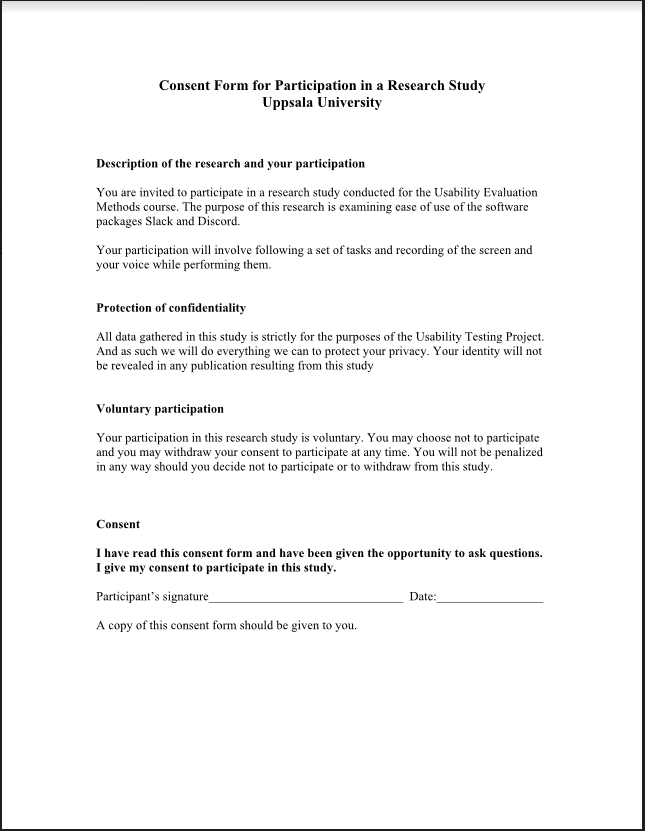
An event is either a state event or a point event depending on if an event is an instant occurrence (point) or if an event occurs for a duration of time (state). The only point event here is the number of WC (wrong clicks) and the rest are state events which are used to determine the duration of the tasks.

# Appendix B: Pre-Test Questionnaire

Before the test begins, subjects participate in a pre-test questionnaire so that their experiences in Discord and Slack may be categorized.



# Appendix C: Consent Form



# Appendix D: Independent Sample T-tests results

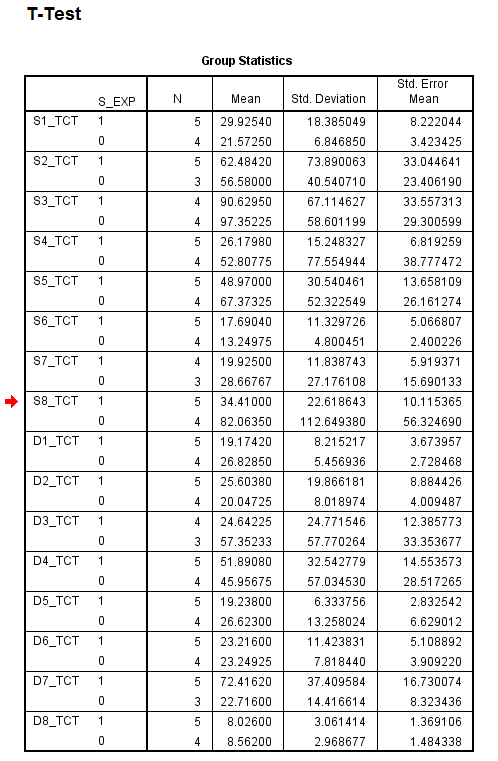


Table Group statistics for individual tasks Grouped by Slack experience

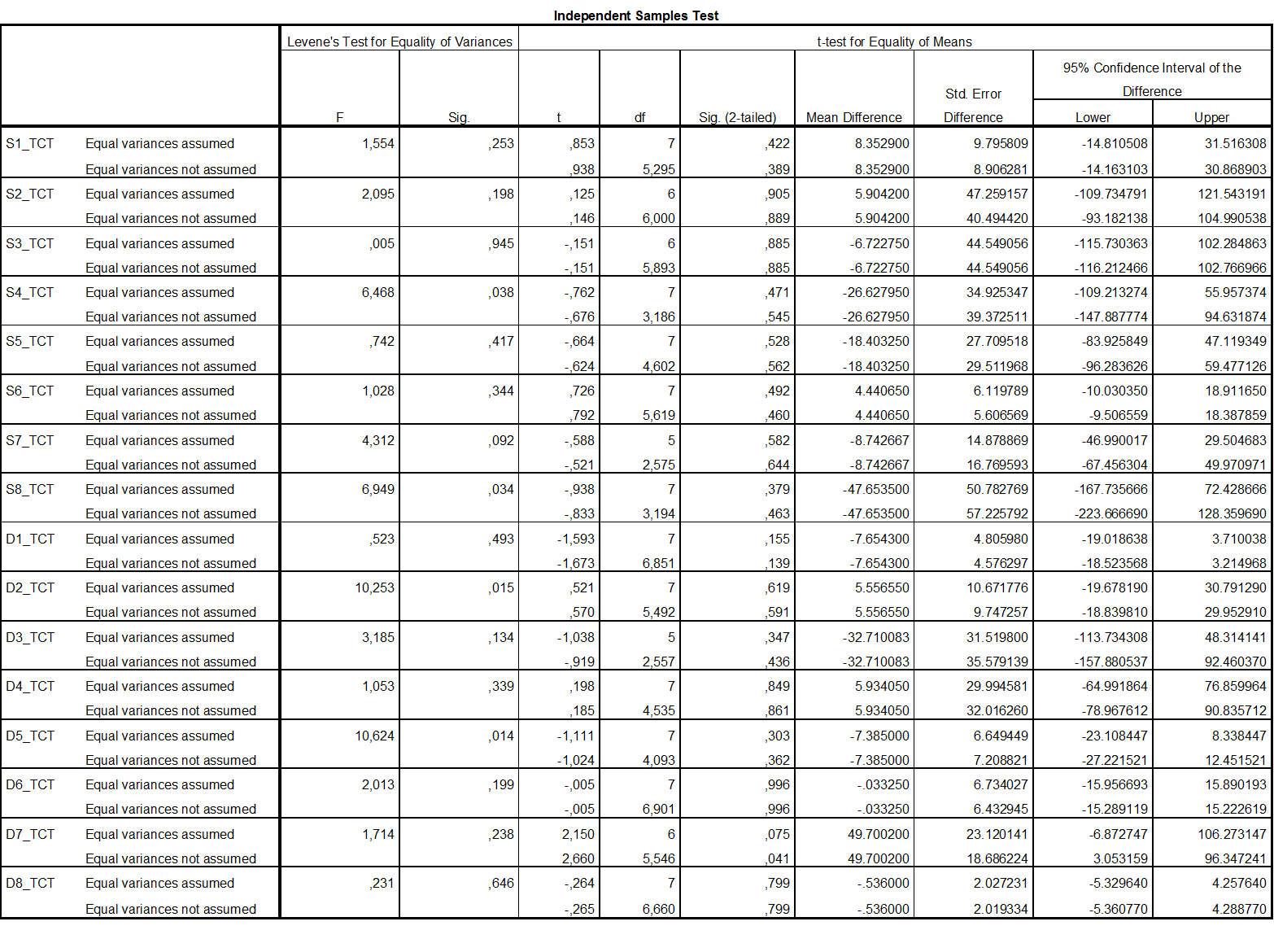


Table Independent sample t-test for individual tasks grouped by Slack experience

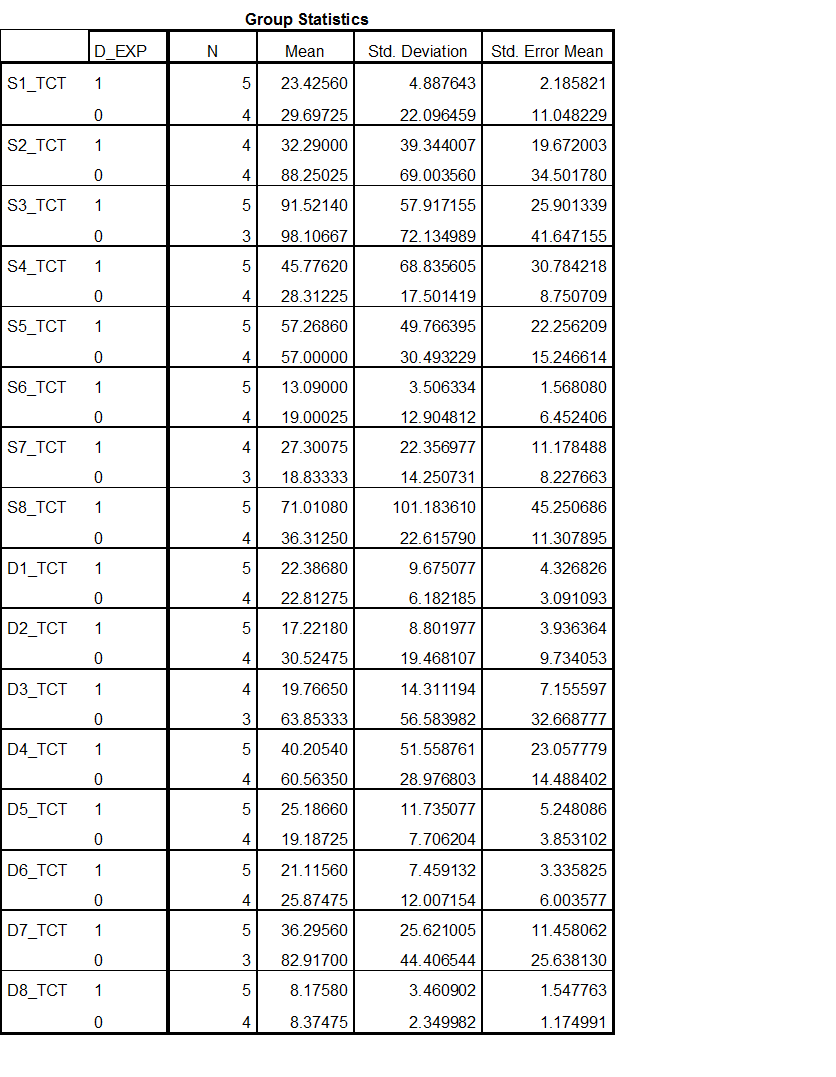


Table group statistics for individual tasks grouped by Discord experience

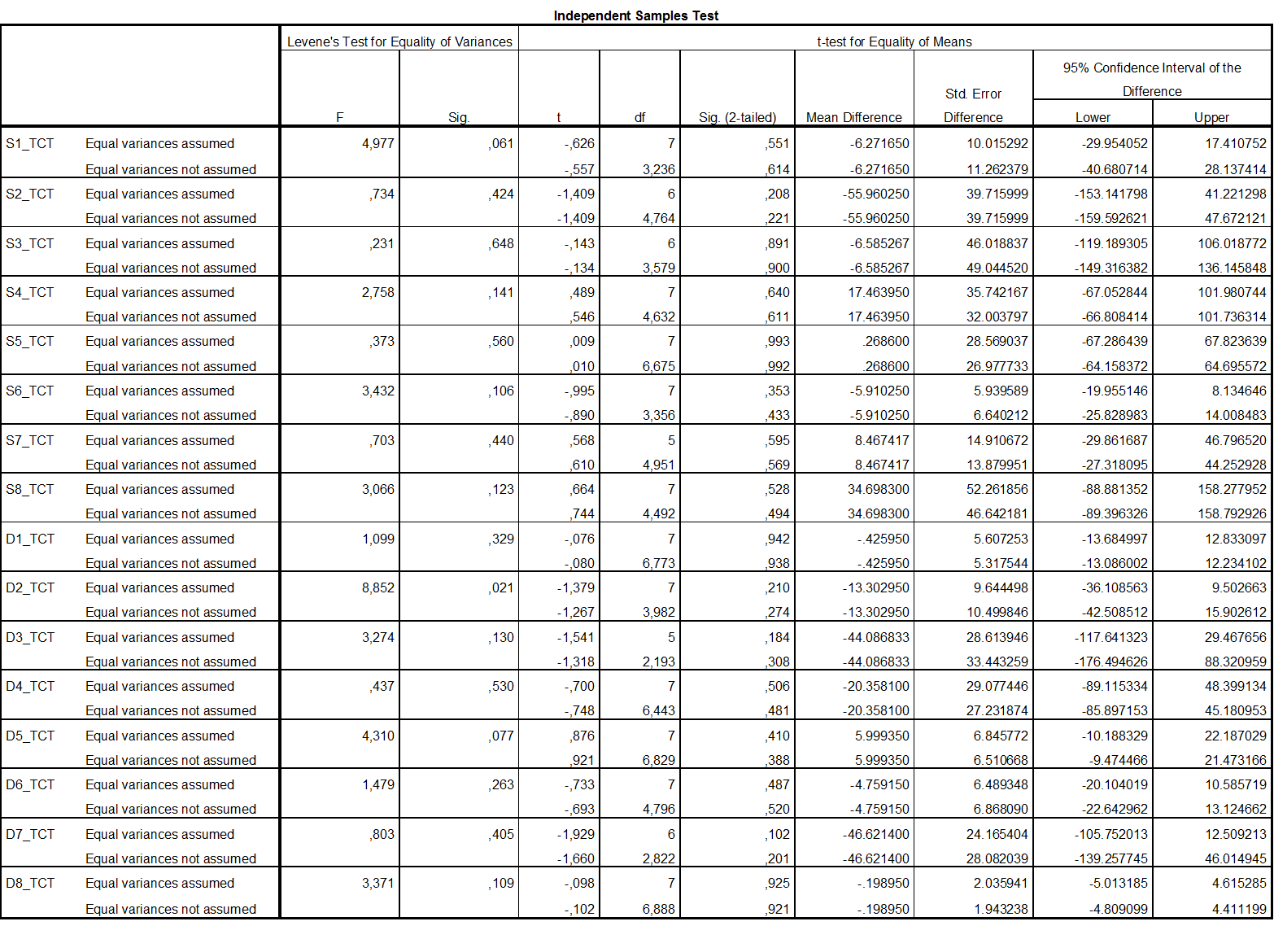


Table Independent sample t-test for individual tasks grouped by Discord experience

Independent sample t-tests for mean task completion time

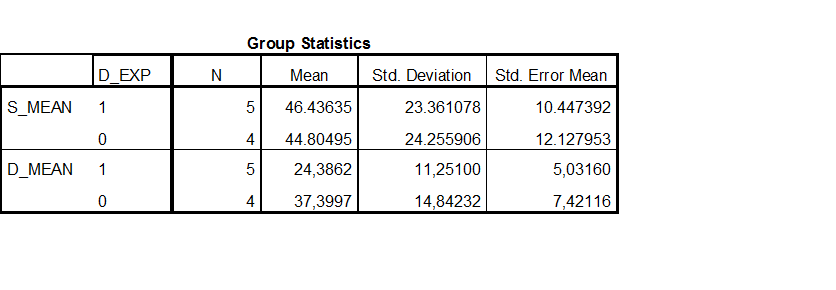


Table Group statistics for mean task completion time grouped by Discord experience

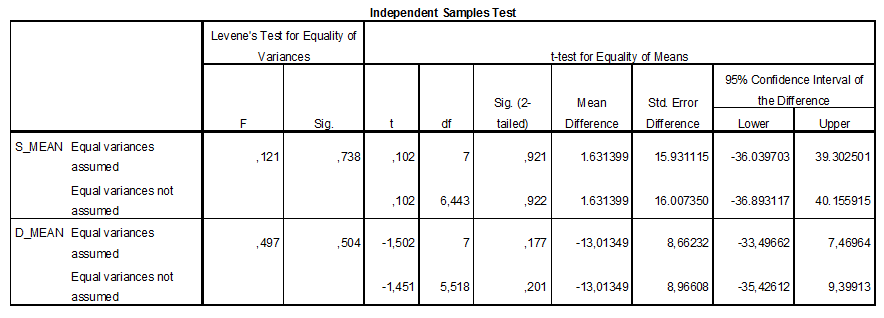


Table Independent sample t-test for mean task completion time grouped by Discord experience

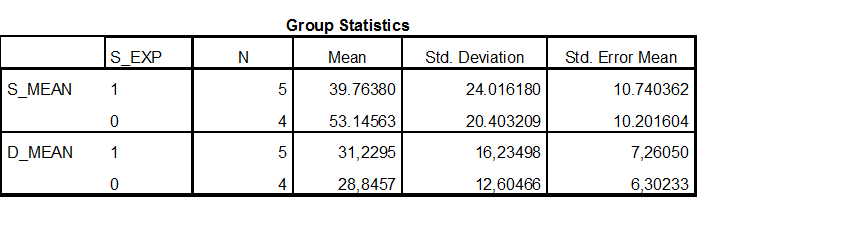


Table Group statistics for mean task completion time grouped by Slack experience

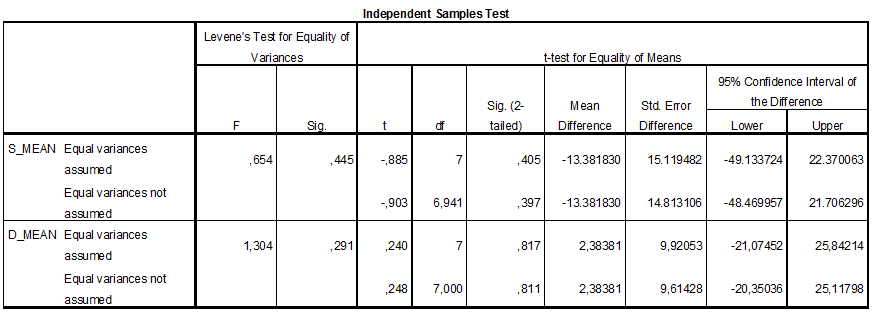


Table Independent sample t-test for mean task completion time grouped by Slack experience