## Usability Report

Given that this game was built from scratch, there’s one important aspect that was not previously researched. A proper usability test was never performed on it thoroughly. Some aspects of the game that were verified rougher, such as movement, may have perhaps had a chance for improvement with time given towards taking user feedback and reworking of those trouble spots.\\

So why is measuring usability important during the trials then? Because ease of use, learnability and repeatability are some of the most important indicators by which satisfaction is evaluated, and the majority of methodologies used in Human-Computer Interaction research to validate assumptions into concrete ideas are variations of usability tests. This present work itself is similar in nature to one method of testing usability, known as Split Testing, whereas two versions of an application are handed to different groups and variations in key metrics are determined. In sum, the measurement of usability is indissociable from any form of research involving user interaction.\\

To perform this study on the system’s usability, we resorted to the System Usability Scale\cite{brooke1996sus}, the SUS. The SUS is an industry standard questionnaire consisting of ten elements to which users respond in a scale between strong disagreement and strong agreement. These questions were adapted to the game into the post-game survey and rearranged, however the function of the SUS’s scoring is intact. Table \ref{tab:Table\_SUS} shows these scores.

|  |  |
| --- | --- |
| Cultural | SUS Score |
| V1 | 87.5 |
| V3 | 80 |
| V5 | 90 |
| V6 | 85 |
| V9 | 90 |
| V11 | 87.5 |
| V13 | 85 |
| V15 | 85 |
| V17 | 92.5 |
| Non-Cultural | SUS Score |
| V2 | 65 |
| V4 | 77.5 |
| V7 | 72.5 |
| V8 | 80 |
| V10 | 80 |
| V12 | 87.5 |
| V14 | 65 |
| V16 | 75 |

Between the two groups it’s apparent that the Non-cultural group had a lesser impression related to the game’s usability. While this was a unanimous complaint, the lowest SUS scores were also coming from the volunteers who most felt the need to verbalize their discontent with the game’s movement. The SUS score ranges from 0 to 100, however these are not percentages spread over a uniform distribution. There’s no single guideline by which all SUS scores are interpreted, but what is usually accepted as a good parameter, is that a desirable threshold to surpass is a SUS score of 68, which represents an approximated average of most SUS questionnaires, or alternatively a score of 80, which is associated with high task completion\cite{sauro2011practical}, although these values may vary depending on the researcher heading the questionnaire, sample sizes and sample selection. There are more benchmarks and percentile breakdowns, but in broad terms, it’s valid to state that the Cultural Group’s is set above the Non-Cultural Group by at least a ranking in its usability , and thus, the choice of gestures has provided a benefit to the Cultural Group.

## Confidence

After the games were completed, the participants were asked about each of the individual tasks they performed on a number of aspects. Aspects such as how they felt about solving them, if the gesture felt natural, among others. One of which, exclusive to the second trial, was their confidence in remembering the task's gestures, and, conversely, if they had trouble remembering it. A few more questions not related to the tasks directly allowed to further get a feel for their personal evaluation on the matter.\\

From these, it was possible to rate each user's global confidence in solving the game's tasks on a scale ranging from 0 to 30. However, this is not a metric that's very useful on its own. Beyond the fact that a user's confidence may be related to personal facets and quirk's, we're also not looking into if the game on its own inspires confidence, but rather if the cultural emblems within a fitting context are a major differentiator for the users otherwise experiencing the same game. This means that merely comparing the Cultural and Non-Cultural Group's Confidence Score wouldn't be enough.\\

It was considered important to go back to the game performance assessment and attempt to categorize the sources of confidence in separate ways. The confidence score would remain a global perspective of the user's evaluation, but a sub-score of it would list out how much of that confidence was misplaced and borne of a user performing the wrong action and still clearing the task. Thus, , we may create two, a participant's Confidence score, and a Misplaced Confidence score. But this approach is still insufficient, as the cultural component is missing. It's by taking in account the types of Gestural Mistakes observed earlier\ref{sec:results\_game\_performance} that this misplaced confidence becomes relevant.\\

Ideally for the thesis' hypotheses, for each user that committed a mistake in the game yet claimed confidence that they solved it correctly, they would have done so because they misremembered a well-fitting cultural emblem in place of their actual mistaken gesture. As such, another score is required to be broken down from confidence, that of the Culturally-Driven Misplaced Confidence.\\

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total | Misplaced | Cultural |
| V1 | 26 | 0 | 0 |
| V3 | 27 | 0 | 0 |
| V6 | 24 | 0 | 0 |
| V9 | 23 | 0 | 0 |
| V11 | 19 | 0 | 0 |
| V13 | 30 | 0 | 0 |
| V15 | 26 | 0 | 0 |
| V17 | 28 | 0 | 0 |
| V2 | 21 | 9 | 8 |
| V4 | 22 | 12 | 8 |
| V7 | 20 | 10 | 10 |
| V8 | 30 | 0 | 0 |
| V10 | 24 | 12 | 10 |
| V12 | 25 | 20 | 20 |
| V14 | 23 | 8 | 4 |
| V16 | 22 | 3 | 3 |

Since the Cultural group made no definite gestural mistakes, their misplaced scores can only be zero, and their culturally-driven misplacement, while possible in practice, would require issues with the task design itself, which was precisely the reason why two were eliminated from the evaluation. One way or another, while confidence is an important metric for the global experience assessment, the whole Cultural group is not really insightful for this current approach. However, as for the Non-Cultural Group, a lot of the score is valid. Since nearly the complete majority of the gestural mistakes were Emblematic Substitutions, it's no surprise that a lot of the confidence the users had misplaced was sourced to their cultural expectations. Manifestly, 85\% of all of the misplaced user confidence score was due to the Non-Cultural trial intentionally breaking their natural conventions. Given an application with more direct forms of feedback where it comes to user failure, this data favors the gesture set employed by the Cultural group enormously.

## Immersiveness Indicators