1. **Project/System Name**
    a. PersonaPal
2. **Project Team**
    a. Ivan Saldarriaga, Wavid Bowman, Tam Huynh, Charis Chen
3. **Application Overview**
    a. The prompt we were given for our project is as follows: A character customization interface that helps users customize the personality or attitude of their chosen character. We started our design by thinking about personas, a set of exemplary attributes of our target audience. We decided, in short, that our target audience is college aged individuals (18-22) that want a simple and easy interface to develop personalized characters using a set of traits. We hope our application can generate personalized and unique characters based on a user's selection of traits, such that they can save and download their generated character for personal use.

    The application is made for computer systems that have a Cuda compatible graphics card. It is possible to run the system on any computer, but the performance will have a sizable increase in character generation time. It is also designed to be used on a computer and not mobile. The following section will go more into detail on our implementation.

4. **Implementation (in depth)**
    a. Our approach to the implementation was to have the user select from several descriptive traits, and generate a script to feed into an image generating AI to develop characters that reflect chosen traits/ attributes.

    We first decided on creating a REACT web app using its framework (HTML/CSS/JavaScript). That comprised all of our front-end languages, all compiled in VSCode. On the other hand, we also created a working backend to save user’s their images, and link with the AI model. This backend was created using python, and stable diffusion was used to generate the character. The standard hugging face SDXL stable diffusion model was used to generate the image, and the system could be changed to generate images from other art styles to meet user needs. Additionally, firebase was used as a backend for object store and user account creation.

    The backend and front end are able to communicate together by hosting the model on a local endpoint using FastAPI. The React application allows users to select variables, which fill in a preset prompt that will be sent over to the backend as a String as a Post Request. The backend then generates the image, and returns a string base8 representation of the image over to the front end in the response body. From there, the front end converts the base8 string back into an image, allowing it to be viewed from the application, and allows saving to the firebase backend as well.

All of the code was done through a public Github repository, where our group members could access and edit the design. Github handles tracking and managing changes to software code as well as storing and allowing for collaboration on software projects. There are API keys listed in the application, but those serve as an example for how to input the API keys and they have already expired.

5. Design Variations

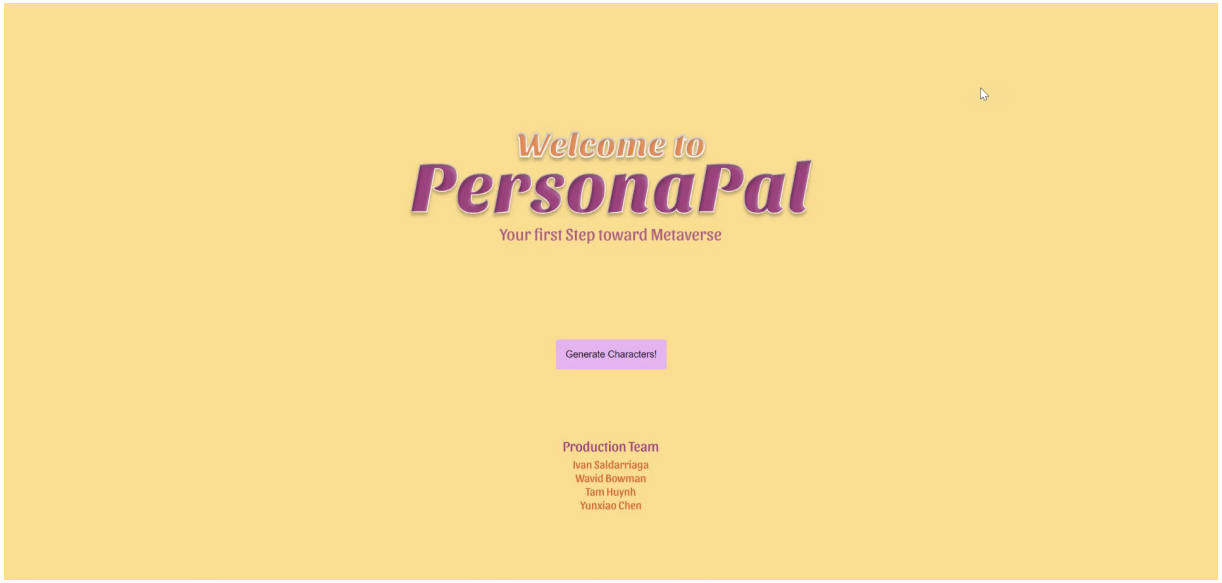
a. Our design variations tested the visual representation of the traits interface and its effect on usability and overall system effectiveness (this is elaborated in the upcoming sections). For both applications, the user starts on a welcome screen that guides them to a landing page (the main interface) where they can start their interaction.

After logging in (using a fake account for privacy concerns) they can start selecting traits. For Design A, the traits are grouped in single-select dropdown menu items with a placeholder text describing the category: time period, vice etc. Within each of those categories, the user can select 1 of the subcategories (Ex. Period has subcategories: modern, futuristic, ancient and medieval). For Design B, the traits have visual indicators (icons) as buttons you can select (same categories and subcategories).

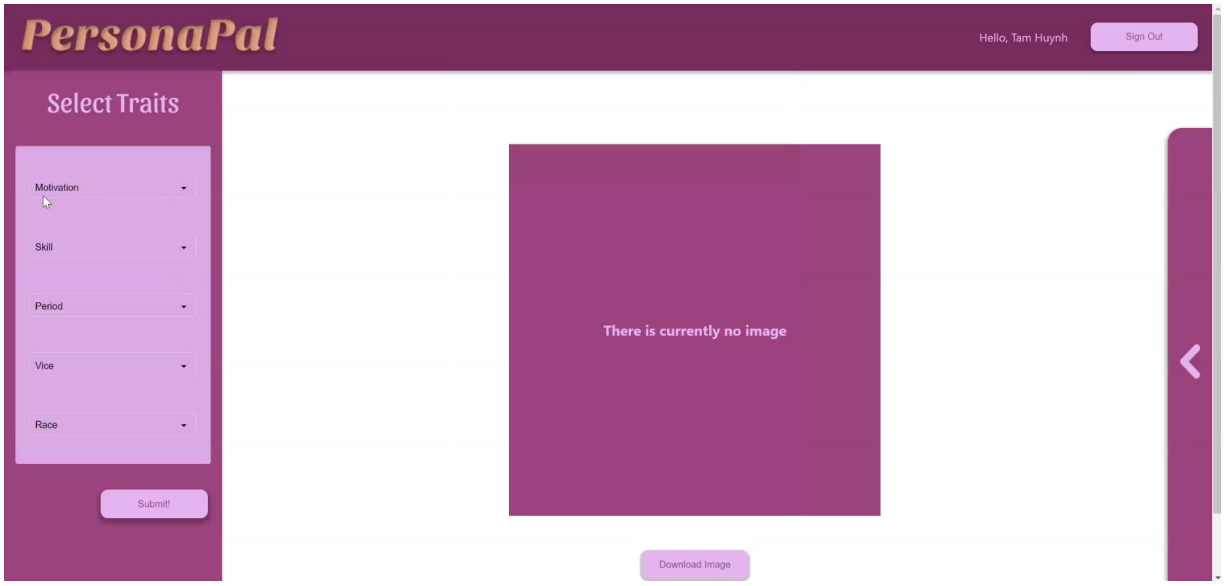
After making all the necessary selections in either design, they can generate an image based on their selections. This takes a few seconds to generate, so they are shown a “Your image is generating…” to communicate that it is loading. Once loaded, and the user is satisfied with their character, they can download (via button) or save (via drawer component on right side) to their account.

If the user returns to the app, they will see that their images that were saved, will still be there for easy access!

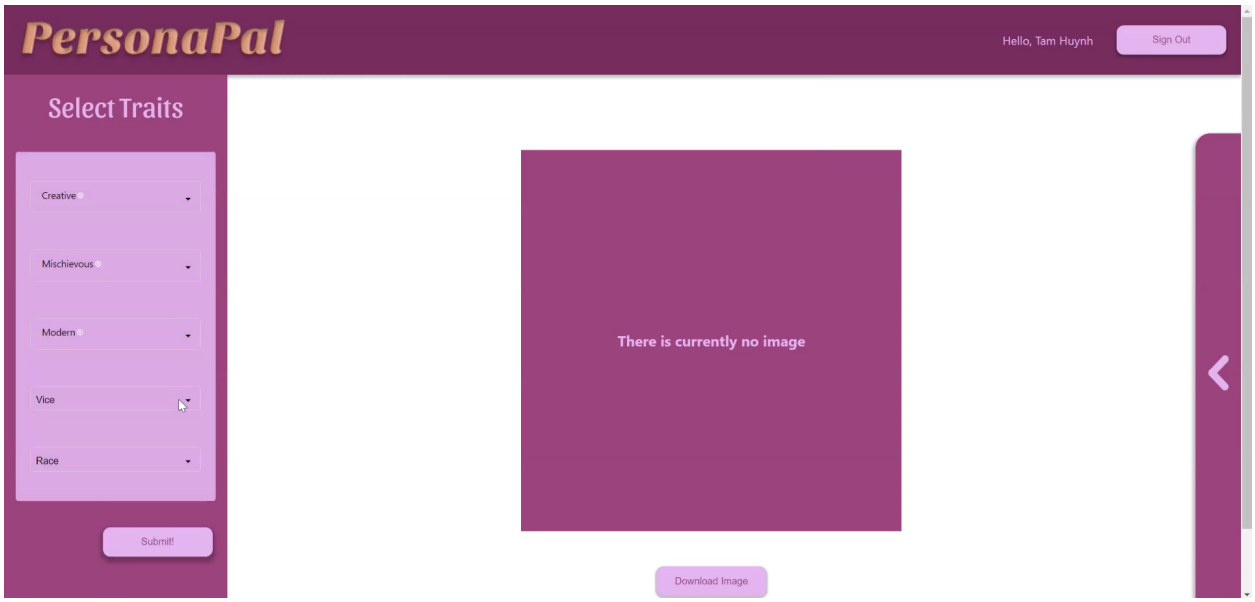
b. Design A:
    i. Reminder of its unique ‘select traits’ component: the traits are grouped in single-select dropdown menu items with a placeholder text describing the category with 4 subcategory options to choose from each.
    ii. Images of Design A implementation:



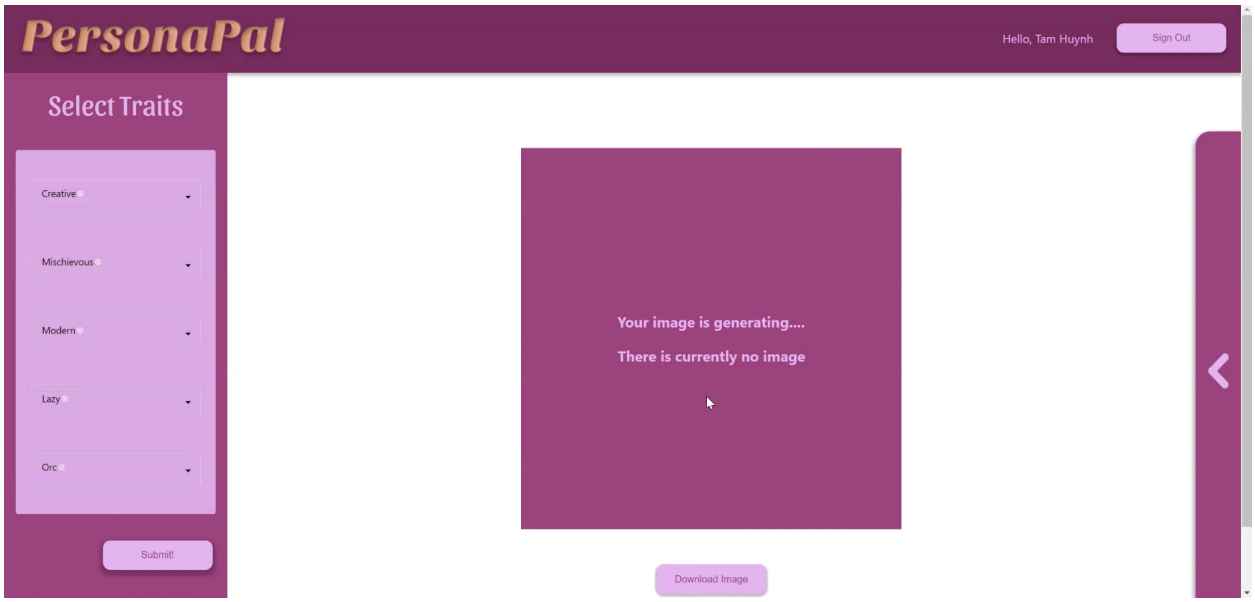
Welcome Screen



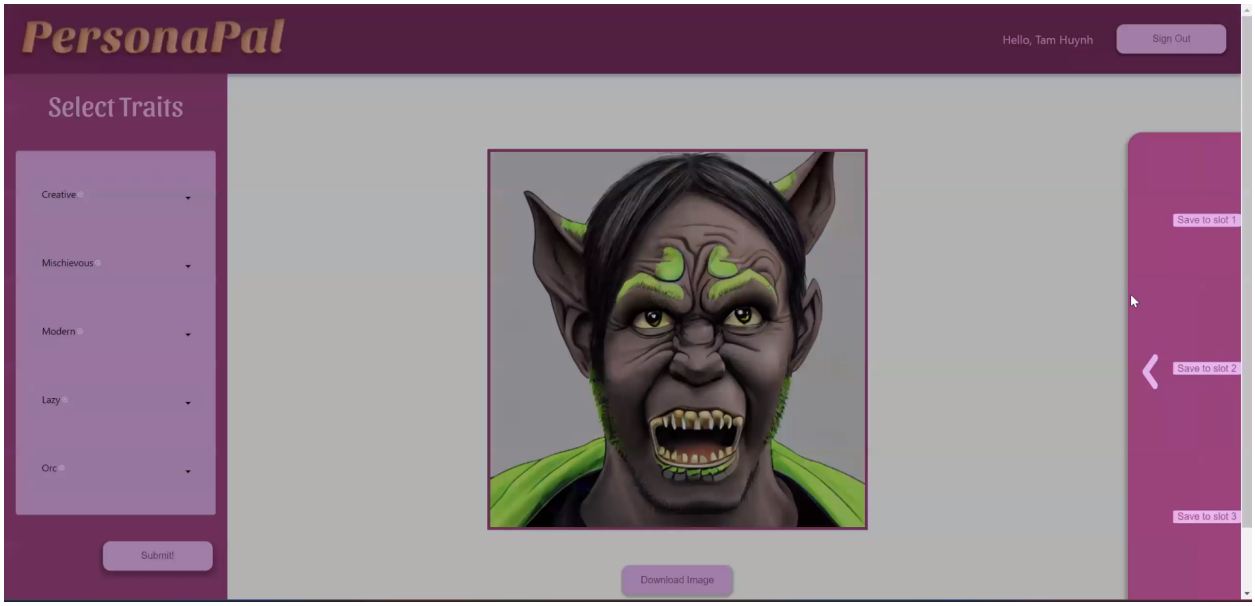
Design A Landing Page



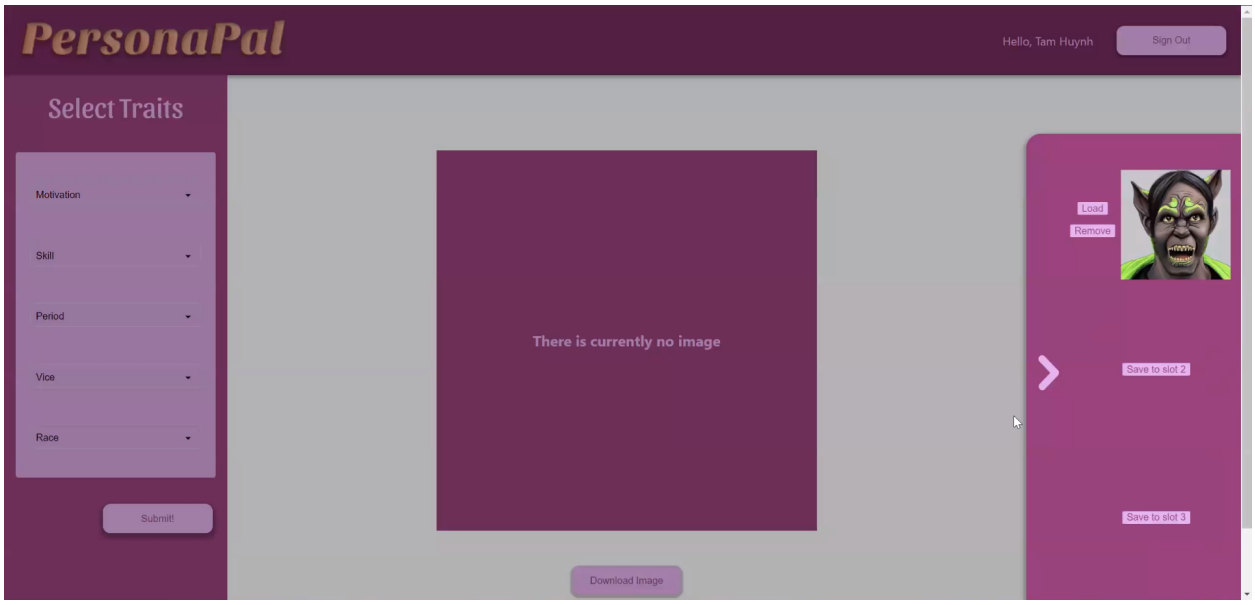
Selecting Traits from Dropdown on Left (Independent Variable)



Communicating Generating Image to User (Middle)

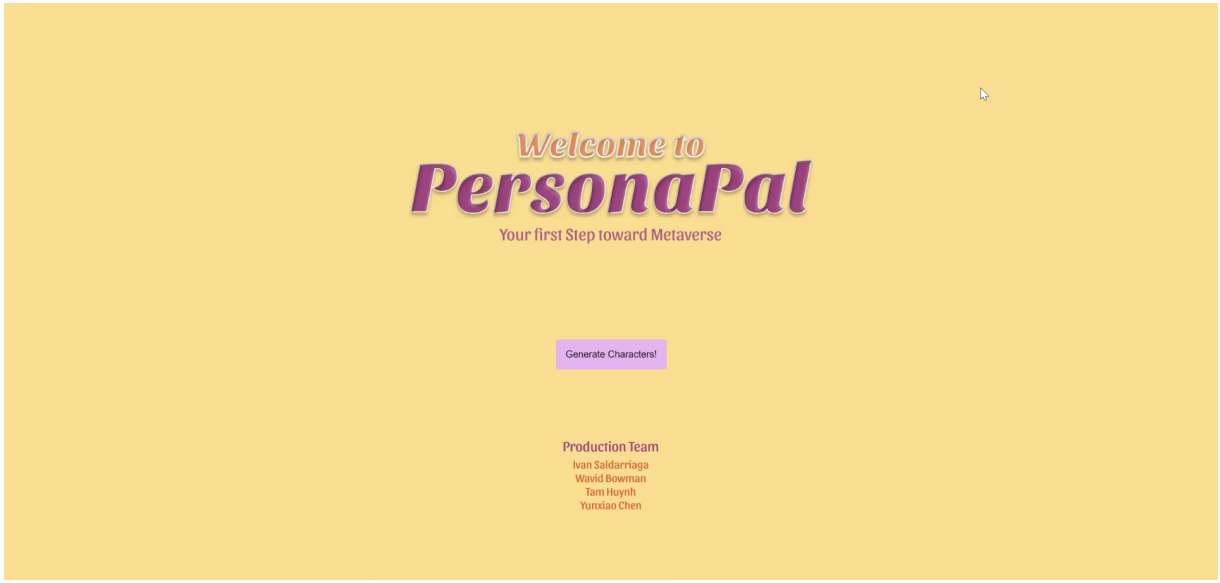


Generated Image

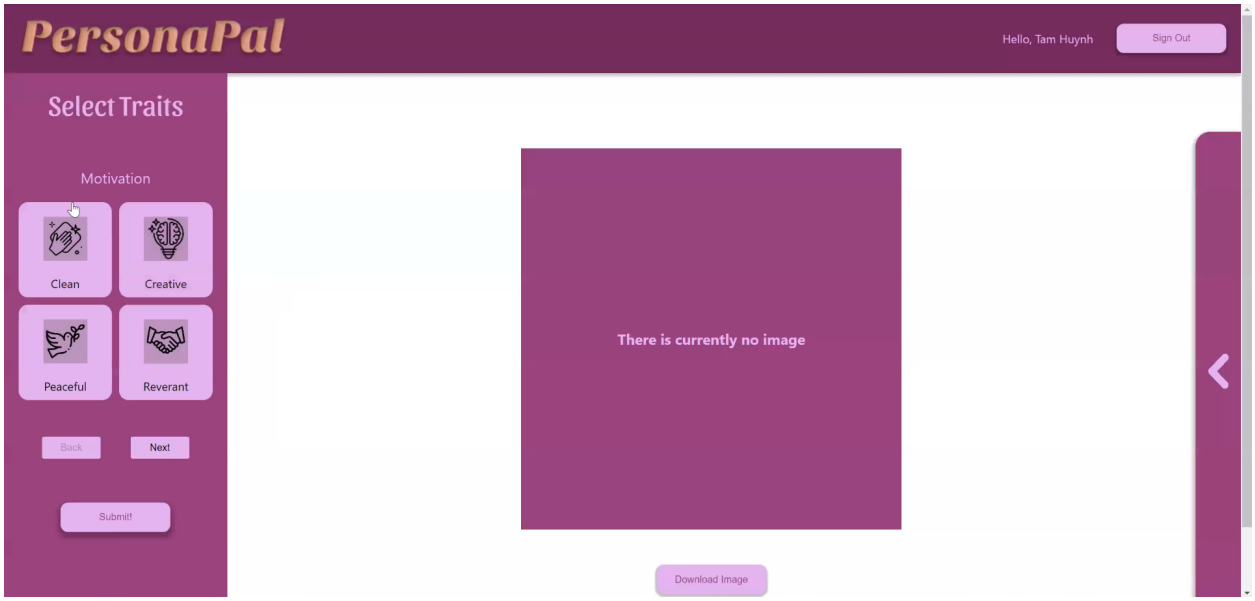


Saving Image (Right)

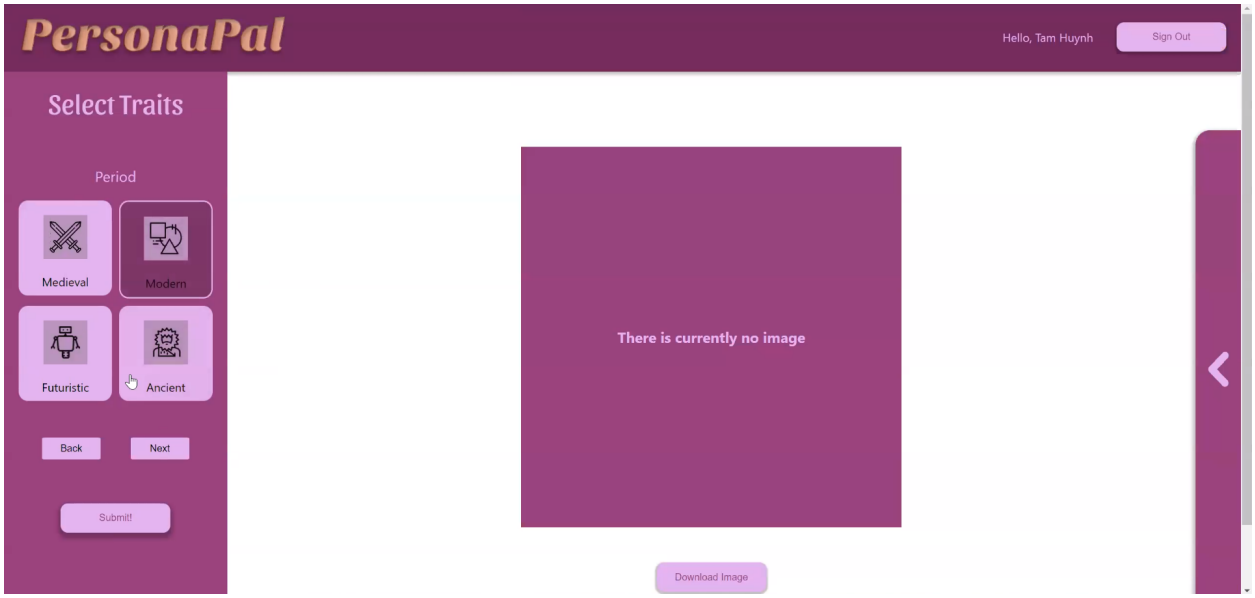
c. Design B:
   i. Reminder of its unique ‘select traits’ component: the traits have visual indicators (icons) as buttons you can select (same categories and subcategories).
   ii. Images of Design B implementation:



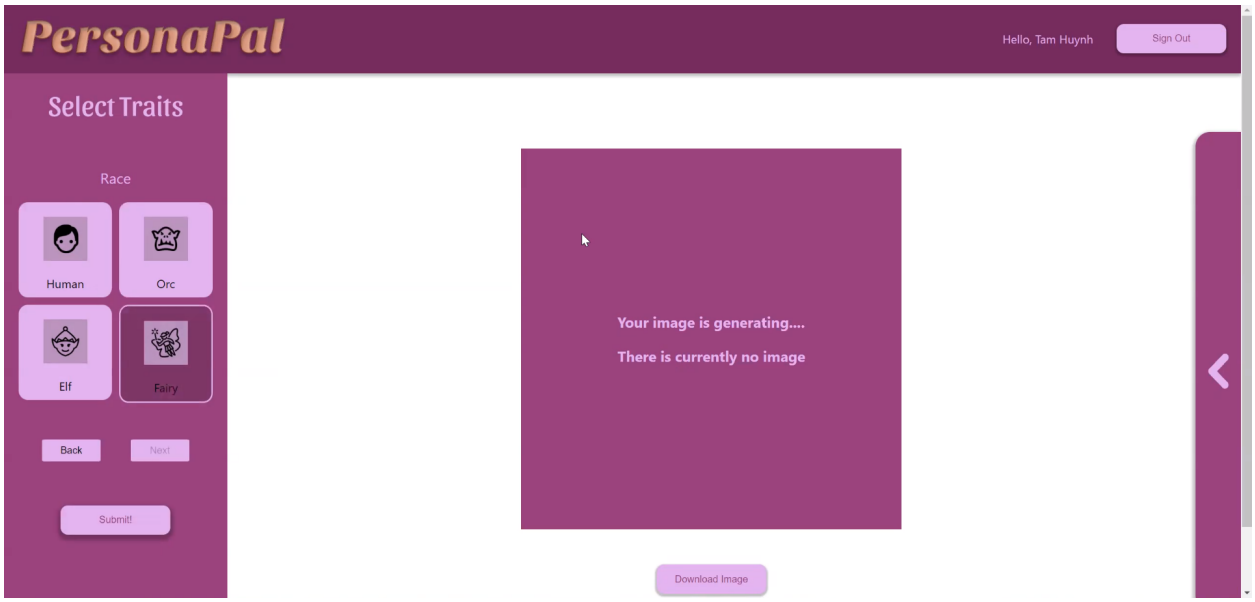
Welcome Screen



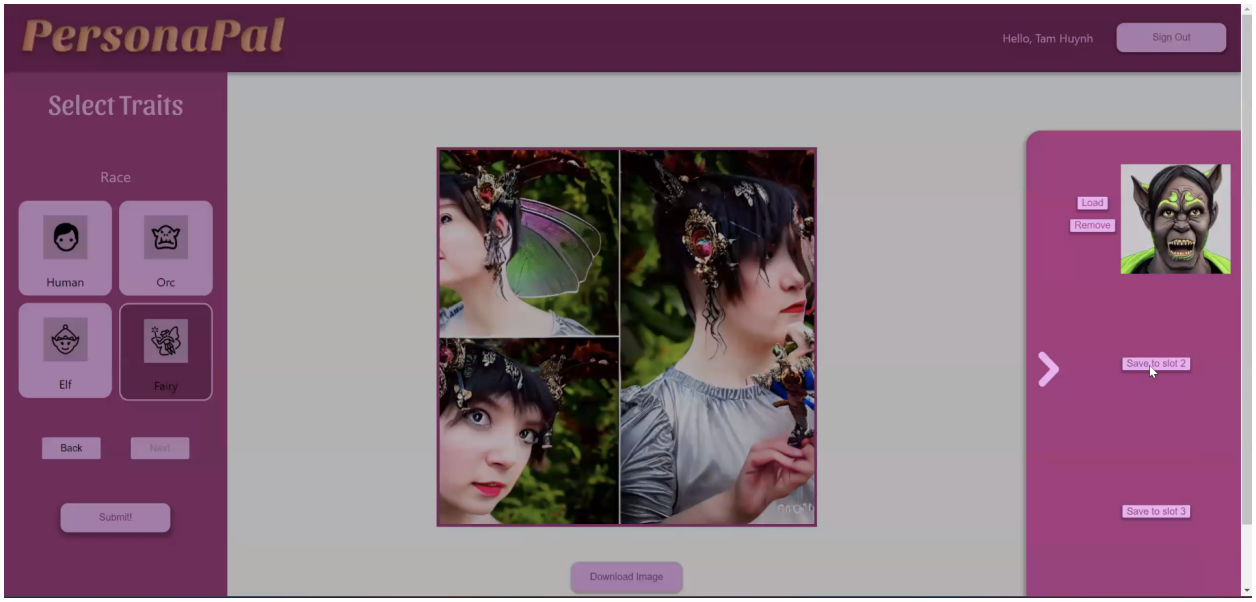
Design B Lading Page



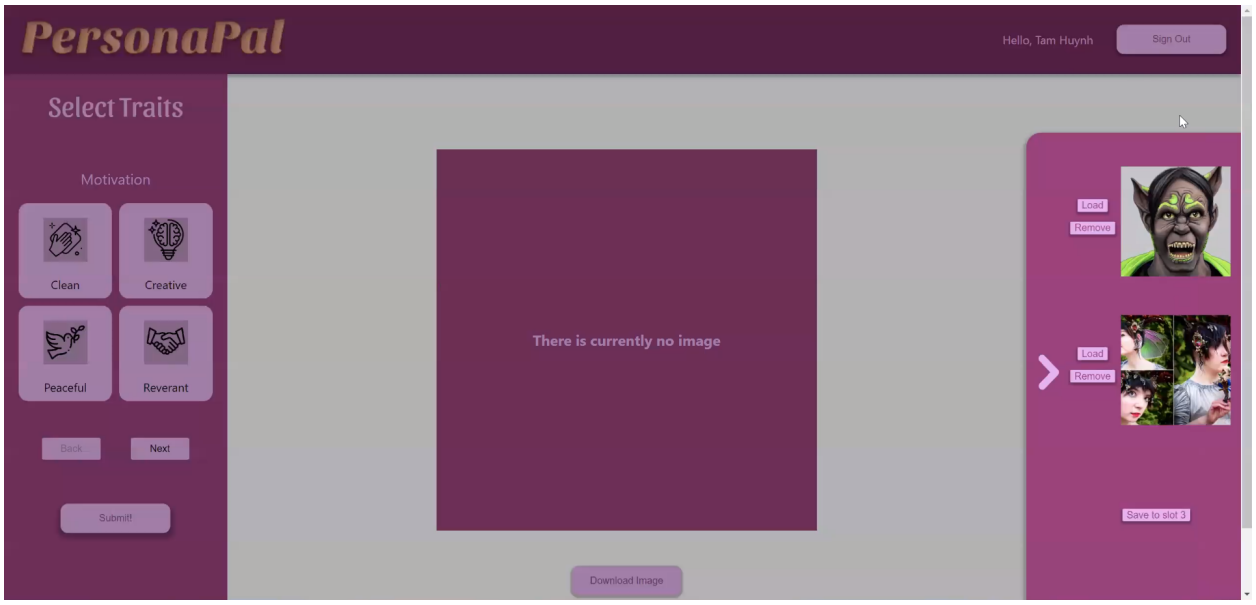
Selecting Traits from Icons on Left (Independent Variable)



Communicating Generating Image to User (Middle)



Generated Image



Saving Image (Right)

6. **User Study 1 (Usability Testing)**
    a. In this experiment, we asked the user to complete 3 separate (but connected) tasks in order to facilitate total use of the system and all of its features. Our specific tasks were as follows: Logging into the application, Creating/ Generating a character, Saving or downloading their character. We conducted this test on 6 participants, some in the classroom as offered, and the rest in a private room to

limit distractions and foster a professional testing environment. Our participants were all undergraduate students at the University of Florida from 18-22 years old.

For this usability test, we asked our participants to use the ‘think aloud’ protocol in order to understand what steps the user was taking to complete each task as well as an open ended question. This would allow us to know what things should be changed for our next suite of tests (AB Test). Aside from the think-aloud transcript taken by a notetaker, we also measured the time it took for each user to complete each task. On each of these metrics, we performed quantitative and qualitative analysis using Rstudio and MaxQDA.

For the quantitative metrics and analysis, we had several self reported metrics from a scale of 1-7 (1 being the poorest performance, and 7 being the highest) that assessed satisfaction, perceived usefulness, perceived ease of use, and aesthetic appeal.
To analyze this data, we decided to calculate the mean and standard deviation to assess how each particular metric performed.

Overall Satisfaction
Question: How do you feel about your overall experience of PersonaPal
- Mean: 5.67
- Standard Deviation: 0.82

Perceived Usefulness
Question: I find this character customization system useful for me
- Mean: 5.5
- Standard Deviation: 1.05

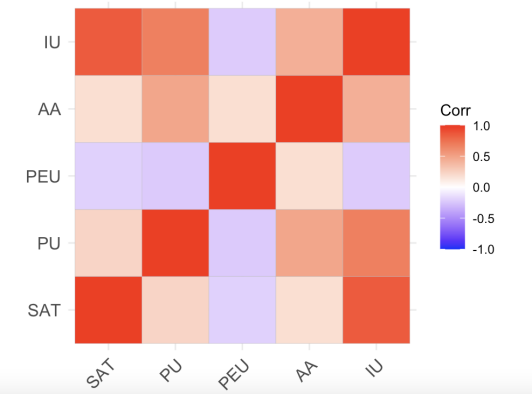
Perceived Ease of Use
Question: How would you rate the ease of use of PersonaPal?
- Mean: 5.83
- Standard Deviation: 0.41

Aesthetic Appeal
Questions: The design of this character customization system is attractive & The appearance of this character customization is visually appealing to me.
- Mean = 5.25
- Standard Deviation: 0.69

To go beyond what was required and to help us gain more insight, a final correlation matrix was executed to obtain the relationships between each variable to understand what metrics worked well together. R is the correlation value, and p is the measure of evidence against the null hypothesis.

The findings suggest that satisfaction and intention of use was positively and strongly correlated ($r = 0.85$, $p < 0.5$).
Satisfaction was not significantly correlated with perceived usefulness ($r = 0.23$, $p > 0.05$), perceived ease of use ($r = -0.2$, $p > 0.05$) or aesthetic appeal ($r = 0.18$, $p > 0.05$).
Intention of use was also not significantly correlated with perceived usefulness ($r = .68$, $p > .05$), perceived ease of use ($r = – .22$, $p > .05$), or aesthetic appeal ($r = .43$, $p > .05$).
We speculate that the correlations stem from the small sample sizes, but still provide general insight on performance and implies that higher levels of satisfaction are associated with higher intentions to our application!

Visual of the Correlation Matrix:



IU= intention to use, AA = aesthetic appeal, PEU = perceived ease of use, PU = perceived usefulness, SAT = satisfaction

The qualitative analysis performed on our think-aloud protocol as well as some comments from our open ended question "*Do you have Any suggestions on improving this character customization system? (the design, the features, or anything else)*" left us with several themes:

1. Enjoyable to Use
    a. Many of the users found the system in the usability test enjoyable to use. Most said generating their character was not difficult. Participant 2 in particular said, "This task [selecting traits and generating a character] was so easy." Two of the users, 2 and 3, made multiple comments throughout their testing about their enjoyment of the color scheme. Specifically, 2 stated, "this is a nice color scheme."
2. Bugs Impair Clarity of Information (select traits dropdowns)
    a. There were two main bugs users stated impaired the quality of information getting to the user: the text color of the dropdown was too light and the dropdown would go off the page when certain series of steps were taken. These were noted in statements like: "Push the dropdowns up some [so they can] fit on [the] page" from

3 and “Change the font color on trait buttons” from 1. These bugs were fixed in all other iterations of design.

3. Sidebar has Unclear Purpose (saving character drawer):
    a. Saving and downloading through a sidebar did not match the mental model of many of our participants. They said things like “it’s not obvious” and “it was confusing at first.” However, 5 did note that it was nice that it could be collapsed down.
4. Login Process is Unorthodox (strange location)
    a. Every participant noted that there was no login option on the landing page of the app, but they had different responses to it. Some, like participants 1 and 2, found it easy to navigate to the app and login from the landing page, while 5 and 6 found it more difficult. 2 states, “there was just one button on the landing page, so I clicked it… logging in after that was easy.” In contrast, 6 said, “there is no indication to go to the next page.”

The study informed us on a few issues with our application design: there were issues with text color within the trait selection, and many users had issues understanding what each trait meant. This is what led us to create ideas for our second design variation that will be in the AB Test, where the comments from Design A specifically were taken into account before conducting the test (ex. background color). Another problem area was the save characters option for some of our participants that were less familiar with similar applications. Although we definitely considered changing its design, we decided on the more pressing issue of the traits selection as our independent variable, as we only required 1 independent variable. A final consideration was done to the login button, which we decided was only an issue because we were required to have a welcome page, it saw little to no change.

This is all reflected in the final themes derived from the qualitative analysis and the lower mean scores on metrics related to aesthetic appeal and usefulness. Similarly, the code for ease of use/simplicity reflects the higher mean scores for metrics such as ease of use and satisfaction. This focus on simplicity is definitely something to take into account for our AB test and future application designs!

7. User Study 2 (A/B Experiment)

a. In this experiment, the independent variable is the visual representation of the select traits menu, with two conditions: (1) Dropdown Text Selections (no visuals) and (2) Image-Based Selections. Participants will be randomly assigned to one of these conditions to assess the impact of the selection component type on overall program usability and effectiveness of the overall system.

Our dependent variables, therefore, are usability of the overall system and effectiveness of the overall system goal. Usability was measured using the

System Usability Scale (SUS) after all tasks to assess the perceived usability of the system as a whole. Effectiveness was measured using the following self-reported metrics:

1. PersonaPal system helped me create a character that met my specific preferences and needs.
2. I was satisfied with the character PersonaPal generated for me
3. I was emotionally invested and engaged in the PersonaPal program

All measured on a 1-5 scale of agreement. These 3 metrics would assess effectiveness in relation to the program and what we believed are the most important components in a good application- enjoyability as a good app is one you would return to.

Each user was asked to perform the same tasks conducted in the usability test, albeit with less guidance and more free reign to explore the app. After pretending to login (using a fake account for privacy concerns), the user selects their traits and generates a character (however many until satisfied) and then saves/downloads their character. Again, these tasks are the same as the usability tasks but with more opportunity for the user to explore and get a holistic opinion on the application.

After the test, we conducted a post-interview including the above mentioned SUS and 3 self-reported metrics as well as a few interview questions for qualitative analysis later on. All self reported metrics were done through a google form, and the interview portion was conducted verbally with a notetaker. The questions asked were to gauge the users opinion on not only the overall application, but also our independent variable in order to gauge opinion on its different variations.

Interview Questions:

1. Were there any specific features or aspects of the interview process that stood out to you as particularly positive or negative?
2. How did you like the 'select traits' interface, did you think its functionality aided or improved your experience creating your persona?
3. Were there any technical issues or challenges you encountered during the interview? If so, please describe them and suggest any improvements or solutions.

It was important to us to ask for any suggestions and especially critiques as the goal is to have as much input to, if wwe were to fully develop this project, create something enjoyable, effective and usable! Hence, our participants followed the following general demographics: college aged students (18-22) with general computer knowledge. We conducted the AB test on 8 participants for each design (16 total). This is far from enough to conduct and retrieve valuable metrics

and data, but with it, we hope to be able to form a good analysis using both qualitative and quantitative analysis tools.

8. User Study 2: Data Analysis

The hypothesis we laid out for the AB tests are as follows:

1. System Usability Hypothesis:
    a. The usability metrics (acquired through SUS) for PersonaPal Design B will see an improvement in usability than that of Design A.
    b. Reasoning: Our usability testing demonstrated that the main issues with design A were an unclear understanding of what each trait meant, and issues with font/text color. With the logo/icon replacing the text, there will be more guidance on what the traits mean and will be more easily readable to the user.
2. Self-Report for Overall System Effectiveness Hypothesis:
    a. System effectiveness (acquired through 3 self reported 1-7 scale metrics) for PersonaPal Design B will see an improvement in effectiveness than that of Design A in terms of each of our 3 metrics describes before (achieving intended goal, satisfaction with character, and personal/emotional engagement)
    b. Reasoning: Our effectiveness metrics, as previously described, is based on overall enjoyability of the application (what we consider to be effective for a character generator). Since there is less confusion with the meaning of what each trait is because of the new icons, the effectiveness will improve. As a bonus, we hope people will enjoy the icons as we believe visuals are more effective at communication than words.

a. Quantitative Data:

We conducted two independent t-tests to compare the mean scores of system usability (SUS) and effectiveness between Group A and Group B. The first dependent variable (DV1), SUS, was derived by calculating the mean value of the 10 items measuring this variable. Similarly, the second dependent variable (DV2) – effectiveness – was represented by the average of three items. Groups A and B are set as the independent Categorical. Descriptives (mean, SD) and the box-and-whisker plots are also included variables. For the R code of this part, please see the file “RCODE_A/B” attached separately.

b. Qualitative Data:
    i. In order to perform qualitative analysis on the responses to the interview questions, we first coded each sentence that we deemed to have significant sentiment to generate a list of level 1 codes. Then, these codes were grouped into level 2 codes, which served as the foundation for the themes reported. For a full list of these codes, please see the file “AB Testing Qualitative Analysis” attached separately.

9. User Study 2: Results

The reinstated hypothesis we laid out for the AB tests are as follows:

1. System Usability Hypothesis:
    a. The usability metrics (acquired through SUS) for PersonaPal Design B will see an improvement in usability than that of Design A.
    b. Reasoning: Our usability testing demonstrated that the main issues with design A were an unclear understanding of what each trait meant, and issues with font/text color. With the logo/icon replacing the text, there will be more guidance on what the traits mean and will be more easily readable to the user.
2. Self-Report for Overall System Effectiveness Hypothesis:
    a. System effectiveness (acquired through 3 self reported 1-7 scale metrics) for PersonaPal Design B will see an improvement in effectiveness than that of Design A in terms of each of our 3 metrics describes before (achieving intended goal, satisfaction with character, and personal/emotional engagement).
    b. Reasoning: Our effectiveness metrics, as previously described, is based on overall enjoyability of the application (what we consider to be effective for a character generator). Since there is less confusion with the meaning of what each trait is because of the new icons, the effectiveness will improve. As a bonus, we hope people will enjoy the icons as we believe visuals are more effective at communication than words.

Reporting of our collected Quantitative and Qualitative data:

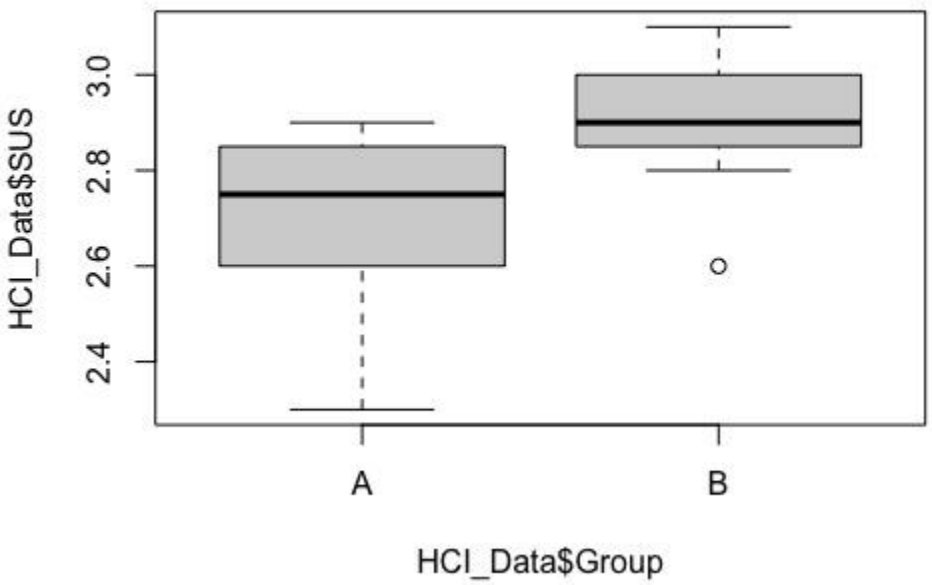
Note: The low number of participants greatly reduces the chances of detecting a statistically significant difference in our results, but for educational purposes, the small sample size will be discounted in our interpretation.

After performing the quantitative analysis using Rstudio by implementing a Welch two sample t-test, the following results were made for each dependent variable measured.

Usability (SUS questionnaire):

Group A:
    Mean: 2.7
    Standard Deviation: 0.2
Group B:
    Mean: 2.9
    Standard Deviation: 0.151
P-value = 0.04187
T-value = -2.2563
Df-value = 13.031

Visualization that shows the Welch two sample test comparisons of the data above:



This means there is a statistically significant difference between both group's since $p < 0.05$. For usability therefore, as we hypothesized an improvement on usability for design B, would reject the null hypothesis that both designs would perform the same. Hence, we can accept our hypothesis that: the usability metrics (acquired through SUS) for PersonaPal Design B will see an improvement in usability than that of Design A.

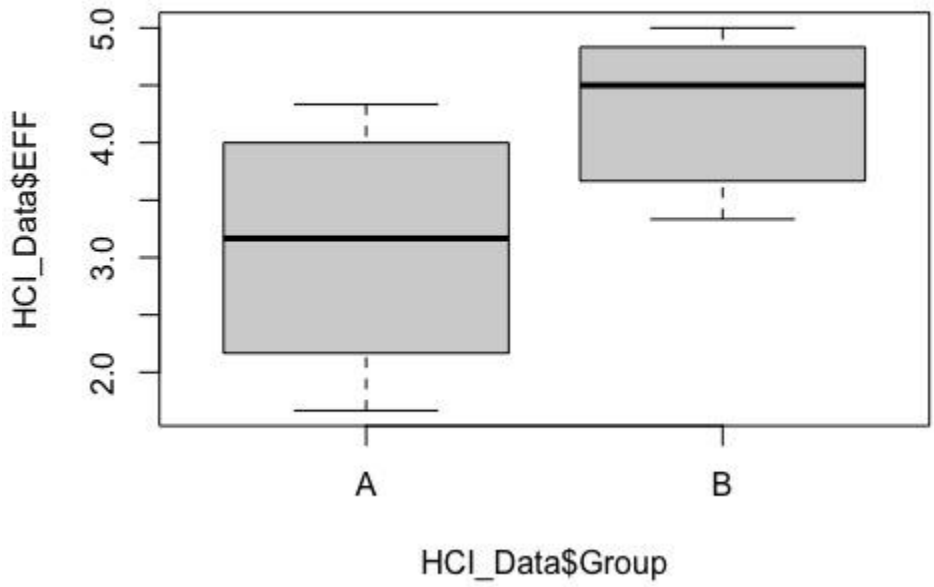
Overall System Effectiveness (3 self-reported metrics):

Group A:
- Mean: 3.08
- Standard Deviation: 1.04

Group B:
- Mean: 4.29
- Standard Deviation: 0.677

P-value = 0.0171
T-value = -2.7632
Df-value = 12.062

Visualization that shows the Welch two sample test comparisons of the data above:



This means there is a statistically significant difference between both group's since $p < 0.05$. For system effectiveness therefore, as we hypothesized an improvement on usability for design B, would reject the null hypothesis that both designs would perform the same. Hence, the evidence allows us to accept our hypothesis that: system effectiveness (acquired through 3 self reported 1-7 scale metrics) for PersonaPal Design B will see an improvement in effectiveness than that of Design A.

After performing the qualitative analysis using MaxQDA the following themes were derived:

1. The user experience is smooth
    a. Many participants stated that throughout the process, in both designs, they found the app easy to use. 2 called the character creation process "pretty simple to use and intuitive;" 3 lauded the simplicity, stating "The simplicity was definitely the biggest thing;" and 5 said, " in the app, it is very easy to use."
    The users also noted that, when discussing technical problems, the process was relatively bugless. 1 stated, "No technical challenges." However, there was a bug with the saved image rendering only after reload, and some noted distaste for the save mechanism, like 7 who said "Like I wouldn't have thought the right

was to save.” No users reported anything breaking or ruining the experience.
Most participants enjoyed the output characters of the app, stating things like “I think it was pretty cool that the model was able to generate images off of the input you provided.”
While these were not variables being controlled for, we are delighted to hear that in both states, the app is simple and easy to use and technically sound.

2. Icons aid visualization, but can be unclear
    a. The consensus of the participants regarding the icons was that it aided their visualization of what the character would look like if a given trait was chosen. This is exemplified by 6’s statement: “It [the icons] made me think more about the actual persona and it was portrayed well on the images.”
    One user noted that it detracted from clarity about how many traits per category users can select, while others stated that it added to the clarity. Specifically, 4 stated, “The icons helped clarify what each selection was.”
    Overall, it affected clarity; clarity improved in some areas, like in visualization and what it will affect, and detracted from clarity regarding the number of traits able to be selected.
3. Dropdowns are easy to use, but hard to visualize
    a. Participants unilaterally stated that the dropdown was easy to use and none stated it was confusing; 6 stated, “I thought it was pretty simple to use.” However, some participants found that using the dropdown made it harder to imagine what the output would look like as compared to using icons. For example, 6 said, “...I didn’t really understand how peaceful was going to be represented,” and 7 said, “I can’t really see how it [the traits] would help make a physical image.”
    Altogether, participants found that dropdowns were very clear on how to use them and select traits, but lacked clarity in how the traits would impact the image; users found that descriptors in word-form alone was not enough to meaningfully visualize how choices would impact the output.

The qualitative analysis provides context to the quantitative results. The application was overall easy to use and the icons from design B helped with the visualization of each theme and subtheme (the overall issue with design A). This reflects and provides further evidence as to why design B out performed design A in terms of both usability and each of the system effectiveness metrics just as we had hypothesized.

10. Discussion

a. To reinstate the results of the quantitative data analysis, the p-value for both usability and each of the system effectiveness came out to where $p < 0.05$. This proves there is a statistical significance to the difference between both designs and the outcome of the results reflect our stated hypotheses. Hence, our null hypotheses (being that both designs would perform the same) can be rejected. Moreover, our qualitative data analysis further demonstrates (albeit subjectively) that design B outperformed design A. This is because the themes derived from design B (the icon/visuals) are overall more positive and improved on the negative themes derived from design A (dropdowns). Since both the quantitative and qualitative data support each other, we can make a reasonable conclusion that the visual representation of the select traits interface (IV from design B) is better than design A in both usability and system effectiveness.

   That is not to say however, that design A did not reflect positive values, relation never fully proves or disproves and hypothesis. In fact, design A got several compliments on its simplicity and ease of understanding. However, we did find it surprising that for effectiveness, design B drastically outperformed design A even with such a small sample size. Both designs were received well in their implementations, yet each suffered their own downfalls. Design A, while it lacked some clarity/ description of each subcategory, was easy to use. Design B, while improved on the visualization, lacked in clarity as reflected in the code descriptions above.

   If there were to be more iterations, we would test other independent variables that could be improved. For example, the sidebar saving option was a common point of confusion for some of our participants. Given more time, we could definitely benefit from conducting an AB test on different implementations of the sidebar/save area.

   As far as guidelines to provide for the community for interfaces regarding character customization, we would suggest that leaning into visual indicators could lead to better interaction. The revelation could suggest this connection is that characters by nature are visual representations of attributes. Having that theme (of visuals) being reflected in the attributes one selects, could provide a better unison and message throughout the entire app!

11. Group Reflection

a. Ivan Saldarriaga Individual Reflection:
   i. I have to admit that when this project began, I failed to see the bigger picture in regards to HCI as a whole. However, after developing and viewing the results of the AB test, it became clear to me how interconnected all the assignments we did in relation to this project were. It was incredible to see at the end how the usability test and its results bled into what we were analyzing for the AB test. Specifically how the

metrics collected from both the qualitative and quantitative data (the issues with the select traits component) ended up being our independent variable to test in the AB test. I felt like I became infinitely more understanding, comfortable and ultimately more appreciative of the intricacies and complexities that go into HCI-related research and discussions. If I were to develop a new project, I would love to experiment more with accessibility and create tests to see how one can make an app, not only accessible, but find metrics that can be universally applied to any project!

b. Tam Huynh Individual Reflection:
    i. When I took UX design in a previous semester, the content I expected to show in class was actually the content that was covered in this class. My understanding of HCI and UX design grew a lot from learning during this semester. I was very naive about how to approach interface design, and I think this class has given me a good grasp on the approach to making them. Specifically, I think I learned the most from conducting the AB-tests and the usability tests. It was a great learning experience conducting those test on an interface I worked on myself, giving me insight on how the process flows for a project I work on and not someone else's. Above all, I've grown an appreciation for the hard work that is put in to make interfaces. While I don't see myself in this line of work in the long run, it was very intriguing to see the process, and I will inevitably work with people in this field down the line so it has been a great learning experience.
c. Wavid Bowman Individual Reflection:
    i. My understanding of HCI was, admittedly, quite low at the start of this project. However, over the course of the semester and over the course of the project, I have gained much confidence in my understanding of the field. Specifically, I enjoyed seeing the change in how well we conducted the early tests like the usability tests to later tests like the AB-tests. After conducting the qualitative analysis on both, I only wish our team had the knowledge we had during the AB tests while we were conducting the first round of tests. Overall, I feel like I have learned a lot about this subject, and, as is the nature of HCI, feel like our project could use another iteration of testing. I hope that for my next experience in a project like this I can see it through multiple iterations of testing and bring to fruition a full product.
d. Charis Chen Individual Reflection:
    i. I've determined to be a user experience researcher more than one year ago. In my own field (communication), we also study HCI with emphasis on theoretical development with less attention to design aspects. To align better with realizing my professional goal, I decided to look into the computer science/engineering domain and also requested to take this class. I was surprised how much work a researcher/designer needs to do

just for improving one or few small features on the product. One key takeaway for me is that the measurements such as perceived ease of use, perceived usefulness, and aesthetic appeal can only reveal a general evaluation of the user experience. That's why qualitative study is very important – we should investigate users' nuanced, complex, and dynamic experience, to find the main point.

e. Final Group Reflection:
    i. All of us had different lasting impressions from the whole process. As a whole, the group has been most appreciative of the interconnectedness of HCI research, whether that be the succession of tests or how qualitative and quantitative analysis work together.

    The most challenging thing about this whole project was finding a rhythm and assessing each individual's strengths and weaknesses. It was a struggle at first, but we got the hang of it, and it was a pleasure to see everyone put their best foot forward for the things that most inspired them. For example, Wavid found strength in qualitative analysis, while Charis spearheaded the quantitative side. Tam was extremely knowledgeable in complex coding structures, and Ivan worked well in managing the team as well as design.

    Each of us found a method to work together, and we were extremely surprised to notice (regardless of sample size) that one of our designs really did outperform the other. Nevertheless, it was extremely rewarding to see how our participants navigated and enjoyed our application and how we could apply the HCI principles learned in class to alter, edit, and improve our design skills!