of an AI system performing half (or more) of the task can remove the sense of proud in an individual. In general, the AI group reported lower levels of negative effect, and for positive affect the results of both groups were closer, with a small advantage for those in the AI group (see Figure 2).

Hypothesis 3 - Increased accuracy

For evaluating the accuracy, we asked the participants to submit their games after done with the task (either by their own or by time-limit). We designed a program to read their game descriptions and attribute points to it. The program contained all the rules required to run the game exactly as required in the document they received to guide them over the experiment. The rules were divided in two sets, *Sprite* rules and *Interaction* rules to match the VGDL game descriptions sets evaluated in this study. One point was attributed for each rule generated corrected by the users. No half points were attributed, or the rule was totally correct or it would not be enough to get a score. Before running the score evaluation, we ran all the games submitted. Those that could not run got a score of zero (0). From the forty-five (45) submissions of the AI group only four could not run. From the forty-two (42) submissions of the noAI group, twelve of them were not running. After this initial stage, we executed the grader software over the entries of the two groups. Then with all the scores available we ran an one-sided t-test to evaluate the hypothesis that the recommender system would increase the accuracy of the participants. The result showed statistical significance ($p \approx 0.004574$). Thirty (30) out of Forty-five (45) submissions from the AI-group got the maximum score (12 points) against twenty (20) out of Forty-two (42) submissions in the noAI group. 67% (AI) against 47% (noAI). In the AI group, excluding the submissions with the maximum scores and the zeroing ones, only two entries got very low scores (2 and 3), most of the others got scores of 8, 9, and 10 points. The noAI group did not have so low scores after excluding zeroing and maximum scores entries, however they could not get close to the ideal (12) and reached average values like 4, 5, and 6 points. In general, by looking at the final results we saw that most of the errors for the two groups (however with more occurrences in the noAI group) were happening in the interaction set. Or they were missing interactions or using the incorrect sprites to missing them, applying interactions that were not required also was a common mistake presented. For our surprise, we saw two entries in the AI group that even provided the termination set of the game, i.e the conditions that define if a gameplay session results in a win or a lose state. Just for the record, no bonus points were given and, of course, the termination set was not used for evaluation in any case.

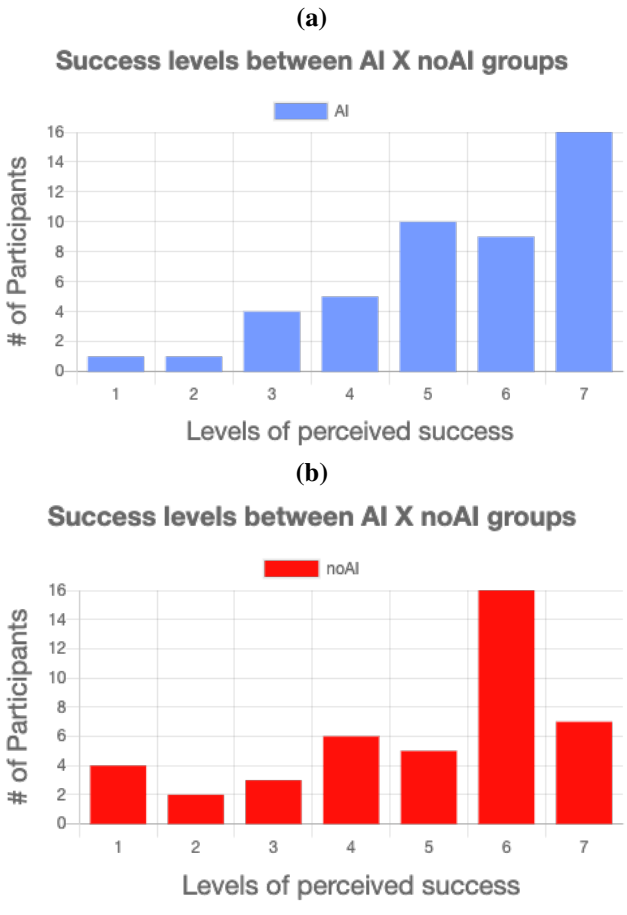

Figure 3: Success levels reported by the respondents in the AI group **(a)** and non-AI group **(b)**.

Hypothesis 4 - Increased Self-Efficacy

We hypothesized that the presence of the recommender system would increase the participant's Self-Efficacy. However, after analyzing that by applying a one-side Wilcoxon-Whitney test, we could not find statistical significance. For all the questions of the computer self-efficacy scale, just one of them reached a value that could be significant ($p \approx 0.047$), however not considered after applying Bonferoni correction.

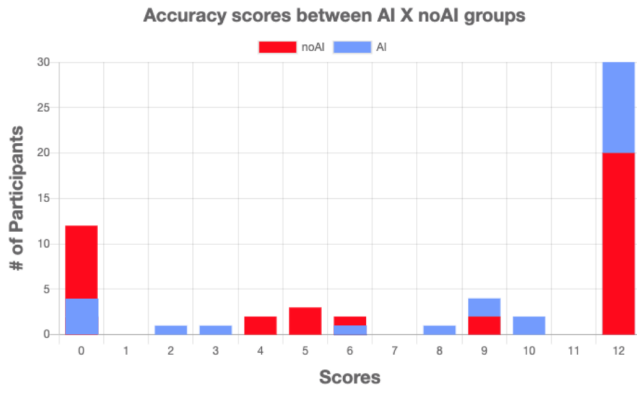

Figure 4: Accuracy scores for the two groups. We can see that users in the AI group performed better than the ones in the noAI group. 30 of the users in the AI group delivered their tasks with maximum accuracy (12 points) against 20 from the noAI group. A complete fail (0 points) was less common in the AI group as well. 4 against 12 total fails from the noAI goup.

The question asked if the participant would be able to perform the task if they never had used a similar product before. Participants' from the AI group were more eager to agree than participants' from the noAI group. This was the only question whose difference of means between the two groups was greater than 1.6. All the others questions had mean difference between 0.27 and 0.82. As the computer self-efficacy scale does not have a method of a evaluate the whole experience as PANAS for example, and because all the questions got results too close for both groups, we do not have how to conclude which of the two groups performed better for this particular (self-efficacy) evaluation.

Qualitative Analysis

We decided to include an open field where the users would be free to report anything they wanted about the whole experience. It was not mandatory and appeared after the forms previously discussed. 40 people submitted their comments, 22 from the noAI group and 18 from the AI one. We categorized them in to effort and positive/negative impact by analyzing their inputs using the online free trial of the Atlas.ti software. We could use this qualitative approach to support our previous findings. Users in the AI group were more positive about the experience, basically their speech are all categorized as of positive impact and low effort. One of the participants stated how easy the whole experience was by saying that *"the game dev kit itself makes it super easy to build a game because it tells you what you can choose and based on what you have chosen, it will tell you the possible interactions that can happen between them."*. Another user stated that the system is self explanatory, *''I really liked how the interactions were suggested so it was self explanatory and leads one to create the game"*. Finally, users also expressed their contentment with the study by affirming that *"it was exciting and fun to play"* and that *"overall it was a fun and pleasant experience"*. In the noAI group, users' answers were more often categorized as indicating negative impact and high effort. Some of them were complaining about the task's time, *"it requires a little bit more time to watch videos and design level as well. I could complete everything except for the level design."*. Others complained about bugs they found in the tool, *"The application broke part of the way through so I couldn't finish the task. I got as far as making the enemies, but the bombs wouldn't go down properly and I couldn't even see the enemy sprite. Then I made some more enemies exactly as the video showed it(down to the sprite), and those wouldn't show up."*. Even when they expressed positive reactions, they were followed by problems they faced during their experience: *"I really liked the UI of the tool but I had a lot of trouble with the interactions."*. In general, we saw that users in the AI group would report better experiences and even enjoyment to some extent because the automatic procedures saved them time and effort to learn and even master UI commands. By contrary, participants in the noAI group had to worry about all the procedures to perform the task since they do not had any kind of automatic assistance.

Conclusion

In this paper we evaluated Pitako, a recommender system for assisting novice game designers, built on the Cicero AI-driven game design assistant. It provides recommendations based on frequent itemset data mining algorithms. Designers get the suggestions while design their games. Their choices tune the system and it is up to them to explore common choices and design clones with small changes, or getting recommendations that lead them to try something new. Because this tool offers components already created and tends to avoid users effort in design everything from scratch, we hypothesized that such a tool would decrease workload (H1), Increase computational affect (H2), Increase accuracy (H3), and finally, increase self-efficacy (H4). We recruited 87 participants and divided them in two groups. We asked them to design the game *Space Invaders*. One group executed the task with Pitako and the other group without it. Our results found with statistical significance that the presence of the recommender system decreases the perceived users' workload, increases their computational affect, and increases their accuracy. No statistical significance was found about the users' self - efficacy. Computer affect is in particular an interesting way to push this work forward. We found statiscal significancy that the participants' in the AI group had a more positive experience as a whole. However, for particular sub-dimensions of computational affect we could not see (with statistical significance) how participants' got influenced by the AI presence. One of them showed that participants' in the noAI group felt more proud in accomplishing the task. Does that mean that the procedural automatic content suggested (or found) by an AI reduces the proudness level of the user? This is still an open question. Participants also gave us their impressions, that we could categorize by analyzing their free-text answers. The AI group reported a more pleasant experience while the noAI group reported their frustration. The presence of a recommender system in the AI group allowed the participants to keep their focus (almost) entirely on the task, while the noAI group had to learn and remind UI commands that exposed them to more mistakes and difficulties in accomplishing the task. We encourage more studies and evaluation of AI-game design assistants with these dimensions in mind: workload, affect, accuracy, and self-efficacy. We are particularly interested in seeing how the experience will change the participants perception when they need to be exposed to the tool for long periods of time, needs to design games of different complexities, and test the tool in both ways (with and without Pitako).

Acknowledgements

To Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), Science without Borders scholarship 202859/2015-0. And to our interns: Katherine LosCalzo, Katalina Park, and ZhongHeng Li.

Provident fuga non ducimus atque pariatur rem voluptates nemo excepturi dolor, distinctio aspernatur dolorum minima fuga, necessitatibus harum atque voluptas neque amet beatae eum quos odio dolorum nulla, perferendis corporis amet veritatis?Quod voluptatibus quas, debitis saepe voluptas molestias dolor voluptates incidunt culpa aliquam placeat alias?Illum sunt delectus perspiciatis corporis ex