# Comparing Player Attention on Procedurally Generated vs. Hand Crafted Sokoban Levels with an Auditory Stroop Test

Joshua Taylor
Thomas D. Parsons
Ian Parberry

Technical Report LARC-2015-02

Laboratory for Recreational Computing
Department of Computer Science & Engineering
University of North Texas
Denton, Texas, USA

February, 2015

# Comparing Player Attention on Procedurally Generated vs. Hand Crafted Sokoban Levels with an Auditory Stroop Test

Joshua Taylor
Department of Computer Science & Engineering
University of North Texas

Thomas D. Parsons
Department of Psychology
University of North Texas

Ian Parberry
Department of Computer Science & Engineering
University of North Texas

## ABSTRACT

Evidence is provided that players pay at least as much attention to a set of procedurally generated Sokoban levels as they do to levels hand crafted by expert designers. Data were collected from 40 participants who played Sokoban under laboratory conditions while simultaneously performing an auditory Stroop test. Three performance measures from the Stroop test were analyzed and compared after accounting for differences in individual players.

## 1. INTRODUCTION

We hypothesize that there is no significant difference between players' attention levels while playing procedurally generated Sokoban levels and their attention levels while playing hand crafted levels. To test this, we assume there is a difference and attempt to measure it using an auditory Stroop test. Under our assumptions, when a player pays more attention to the game and less to the Stroop test their reaction time will slow down, their accuracy will fall, and the number of times they fail to respond will increase.

Sokoban is a grid-based transport puzzle. The goal is to push boxes onto marked goal squares using the player's avatar (see Figure 1). The challenge comes from the placement of the walls, goals and boxes and the restriction that the avatar can only push boxes, and then only one box at a time. For example, Figure 3 shows a level with a single box and a single goal that can be solved in 9 moves: push right, move down, move right, push up, push up, move right, move up, push left, and push left.

Culberson [4] has shown that Sokoban is PSPACE-complete, meaning that it is in a sense at least as difficult as almost any single-player game (Demaine [5]). This, together with its simple rules, makes Sokoban a challenging candidate for procedural generation of interesting puzzle instances of varying levels of difficulty. Some research into what makes a Sokoban level interesting and what makes it difficult include Ashlock and Schonfeld [3], and Jarušek and Pelánek [9].

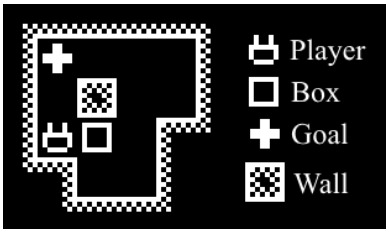


**Figure 1: A simple Sokoban level on the left, with the icons on the right representing the player avatar, a box, the goal, and a wall.**

Doran and Parberry [6] suggest five criteria for successful content generation: (1) *novelty*: contains an element of randomness and unpredictability, (2) *structure*: is not merely random noise, but contains larger structures, (3) *speed*: can be quickly generated, (4) *controllability*: can be generated according to a set of natural designer-centric parameters, and (5) *interest*: has a combination of randomness and structure that players find engaging. Taylor and Parberry [20] presented a procedural Sokoban level generator and argued that it satisfies the first four of these criteria, leaving the topic of player interest for future work. That is the topic of this paper.

We performed a study involving 40 participants who played Sokoban under laboratory conditions while simultaneously performing an auditory Stroop test. While the participants were playing the game, we measured their attention as an indicator of their interest and engagement. Specifically, we measured the attention of participants playing Sokoban levels from the procedural generator of Taylor and Parberry [20] and compared the results to their attention levels while play-

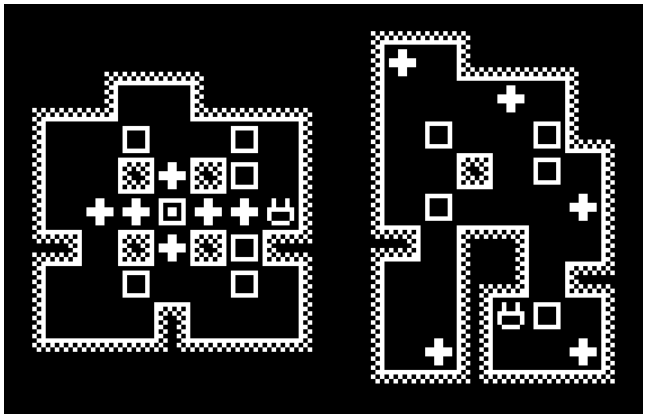

**Figure 2: A human generated Sokoban level on the left, and a procedurally generated level on the right.**

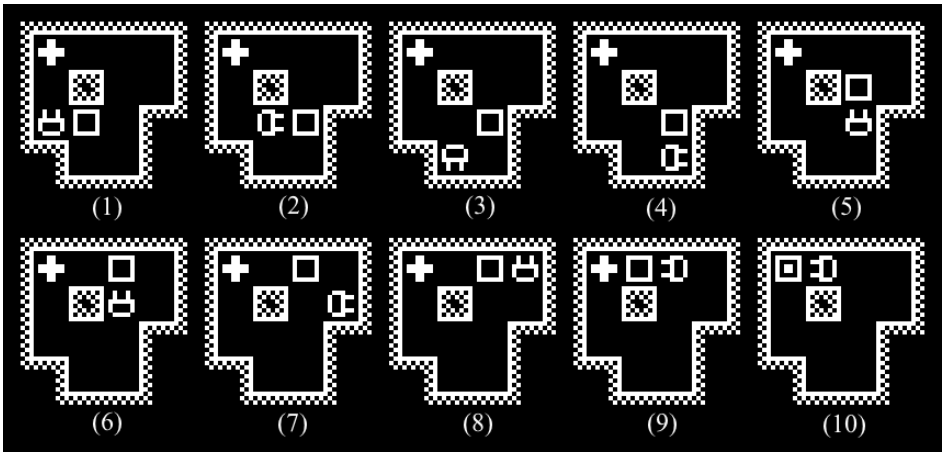


**Figure 3: Solving a simple Sokoban level in 9 moves from the initial configuration (1).**

ing hand crafted levels from experienced Sokoban designers. Figure 2 shows an example of a hand crafted level on the left and a procedurally generated level from our algorithm on the right.

We analyzed the results with three linear mixed models with all dependent variables and covariates modeled as fixed variables and the subject as a random variable. Our results showed little, if any significant difference in player attention between the two types of levels; therefore, we conclude that our procedurally generated Sokoban levels are at least as interesting and engaging to players as human designed levels.

The main part of this paper is divided into three sections. Section 2 describes our study in more detail, including experimental design, analysis issues arising from the data distribution, the effects of boredom and difficulty on attention, and our choice of covariates. Section 3 describes the data gathered, discusses some properties of it and of the models used in our analysis. Section 4 contains our analysis of the data, including analyses of player reaction times, percentage of correct answers, and percentage of unanswered questions from the auditory Stroop test.

## 2. THE STUDY

Attention is a finite resource in the sense that the more attention you pay to one thing, the less you have to spend on other things (Sinnett *et al.* [18]). We measure attention by requiring participants to play Sokoban while simultaneously taking a *Stroop test*, which is a common way of measuring subjects' reactions to conflicting information (Stroop [19], MacLeod [13]). In its original form, words such as "Red", "Green", and "Blue" are displayed in different colors that do not necessarily match the words themselves. Participants are required to respond to the color of the text while ignoring what the text actually says.

Since Sokoban is primarily a visual game, we chose an auditory Stroop test to minimize the direct disruption to play. Participants played Sokoban with their left hand using the W, A, S and D letter keys while simultaneously responding to the Stroop test with their right hand using the 8 and 2 keys on the numeric keypad. The participants played Sokoban without audio, while the Stroop test had no visual representation. While it is easier to multitask across different sensory modalities, there is often still some loss of performance (Sinnett *et al.* [18]). For this experiment, the Stroop test chosen involved a voice saying the word "High" or "Low" in either a high or low pitch. Participants needed to respond to the actual pitch of the word, and not the word itself. For example, if the participant hears the word "High" in a low-pitched voice, the correct response is "Low". MacLeod [13] surveyed various forms of auditory Stroop tests that have been studied, including tests using high and low pitches. The study concluded that auditory Stroop tests may not be as effective as the original but they are similar.

Task engagement involves a user focusing on information that is relevant to a given task and the filtering of information that might interfere with it. In the auditory Stroop task, for example, the impact of an incongruent word needs to be controlled, which effectively renders it salient. This effect has been found to be counteracted by increasing the saliency of the task-relevant input. For example, Krebs *et al.* [12] assessed the influence of novelty on interference processing. They employed a picture-word interference task in which they manipulated the novelty of the task-relevant picture. They found that picture novelty reduced typical Stroop interference from incongruent words. Similar findings were found by Armstrong *et al.* [2] when they presented users with Stroop stimuli while they played an action video game. They found that the executive network was activated during low-engagement gaming conditions and a salience network was activated in response to highly engaging gaming conditions. Therefore, we make the assumption that measuring responses to a Stroop test will allow us to say something meaningful about the subject's engagement.

Stimuli used in the auditory Stroop test were words ("High", "Low") presented in a high and low pitch and were recorded in a high-quality digital format (sampling rate = 44,100 Hz). Exemplars of each stimulus were generated using text-to-speech software to provide a consistent voice, generating

stimuli with similar durations (249 to 275 milliseconds), and pitch (high = 222 Hz, low = 124 Hz). Different pitch-word combinations produced congruent and incongruent conditions. Spoken words were presented via headphones. The speech was clearly audible above the background computer lab noise. For each auditory trials, participants discriminated the pitch by pressing one of two buttons. In-house software was used for stimulus presentation and response logging. Responses were recorded until the next stimulus was presented.

We measured performance on the Stroop test in three ways: (1) *reaction time*: the time between when the word was said and the time when the participant pressed a button in response, (2) *percentage correct*: how often they responded correctly, and (3) *percentage unanswered*: how often they failed to respond at all. For all three, the results were separated into responses to *congruent tests* (those where the word and the pitch matched) and *incongruent tests* (those where they did not match). A reasonable expectation is that as participants pay more attention to the Sokoban game, they will take longer to respond to the Stroop test, their percentage of correct responses will go down, and their percentage of unanswered prompts will go up.

The remainder of this section is divided into four subsections. Section 2.1 describes our experimental design. Section 2.2 discusses some analysis issues arising from the data distribution. Section 2.3 discusses the effects of boredom and difficulty on attention, and our selection of covariates related to this.

## 2.1 Experimental Design

Each participant performed six sub-tasks during the experiment: take the attention test, practice the Stroop test, practice playing Sokoban, play the hand crafted levels, play the procedurally generated levels, and take the demographics survey. Half of the participants took the attention test before the game and half took it afterwards. Half of them practiced the Stroop test first and half practiced Sokoban first. Half played the hand crafted level set first and half played the procedurally generated levels first. Finally, there were five different sets of hand crafted levels from five different authors, chosen for their similarity. Everyone played both practice rounds before playing the main game and everyone took the survey as the last step. This gave a total of 40 combinations of conditions and each subject was randomly assigned to a different combination.

The entire process took place in a single two-hour session per person. The attention test chosen was the Attention Network Test (see Fan *et al.* [7] and MacLeod *et al.* [14]), which is a standard attentional test in psychology research. The demographics survey was based on the 2010 U.S. census with some additional questions about the participants' use of computers, game playing habits and preferences (for example, what types of games they enjoyed), and about their experiences with the study.

The participants were a mix of computer science and psychology students. They were not informed of the purpose of the test, nor that some of the levels were procedurally generated. None of the participants mentioned any differences they may have noticed, either in the comments on the survey or to the researchers directly.

## 2.2 Normality of Dependent Variables

Response time in general does not appear to follow any simple distribution, although some research suggests that a combination of an inverse Gaussian distribution and an exponential distribution is a good approximation (Schwartz [17]). Due to the skewness of the data and the fact that reaction time cannot be negative, we also tried transforming the data. We tried a Box-Cox transformation as suggested by Osborne [15], but since the transformation parameter was close to 0 (approximately -0.35) we chose to use a log transformation for its ease of interpretation.

A *quantile-quantile plot*, abbreviated *QQ plot*, is a graphical device used to check the validity of an assumed distribution for a given data set. Comparing the Normal-QQ plots (see Wang and Bushman [22]) of the congruent and incongruent reaction times in Figure 4, we see that the log times are close to normally distributed and have fewer outliers. For the remaining analysis we use the log response times.

As counts out of a set of attempts, accuracy and "no response" percentages are binomial instead of normal. Those variables were transformed using the logit transformation (see Altman [1]) to compare the log odds ratios.

While many statistical methods assume normality, most of those same tests are considered robust to violations of that assumption. Some research suggests that while normality may not be necessary, such normalizing transformations can improve the power of such methods (Altman [1] and Kirisci [11]).

## 2.3 Boredom, Difficulty and Other Covariates

Regardless of whether or not there is a difference between the hand crafted and procedurally generated levels, it is reasonable to expect that boredom should have a significant effect on attention. The test as a whole was two hours long and several subjects commented that it was getting boring by the end. To compensate for the possible effects of boredom, the two possible orders of play were balanced against each other. Assuming there is no strong interaction between the order of play and which level set was being played, any order effects should average out.

Difficulty is a hard thing to estimate mechanically (Jarušek and Pelánek [8]). Despite choosing levels that were numerically similar to one another (similar size, similar number of boxes, similar number of moves needed to solve, etc.), it seems like the hand crafted levels were still much harder than the procedurally generated levels. Subjects successfully completed many more procedurally generated levels within the same time frame ($1.5 \pm 0.6$ hand crafted levels versus $5 \pm 1$ procedurally generated levels, on average). At the same time, difficulty does not necessarily correlate with interest since different individuals have different preferences for difficulty. (This is supported by evidence from the participant survey. See Section 3.2.) To ensure the differences in difficulty influenced the results as little as possible, several covariates related to difficulty were chosen. These were the number of levels attempted during each level set, the num-

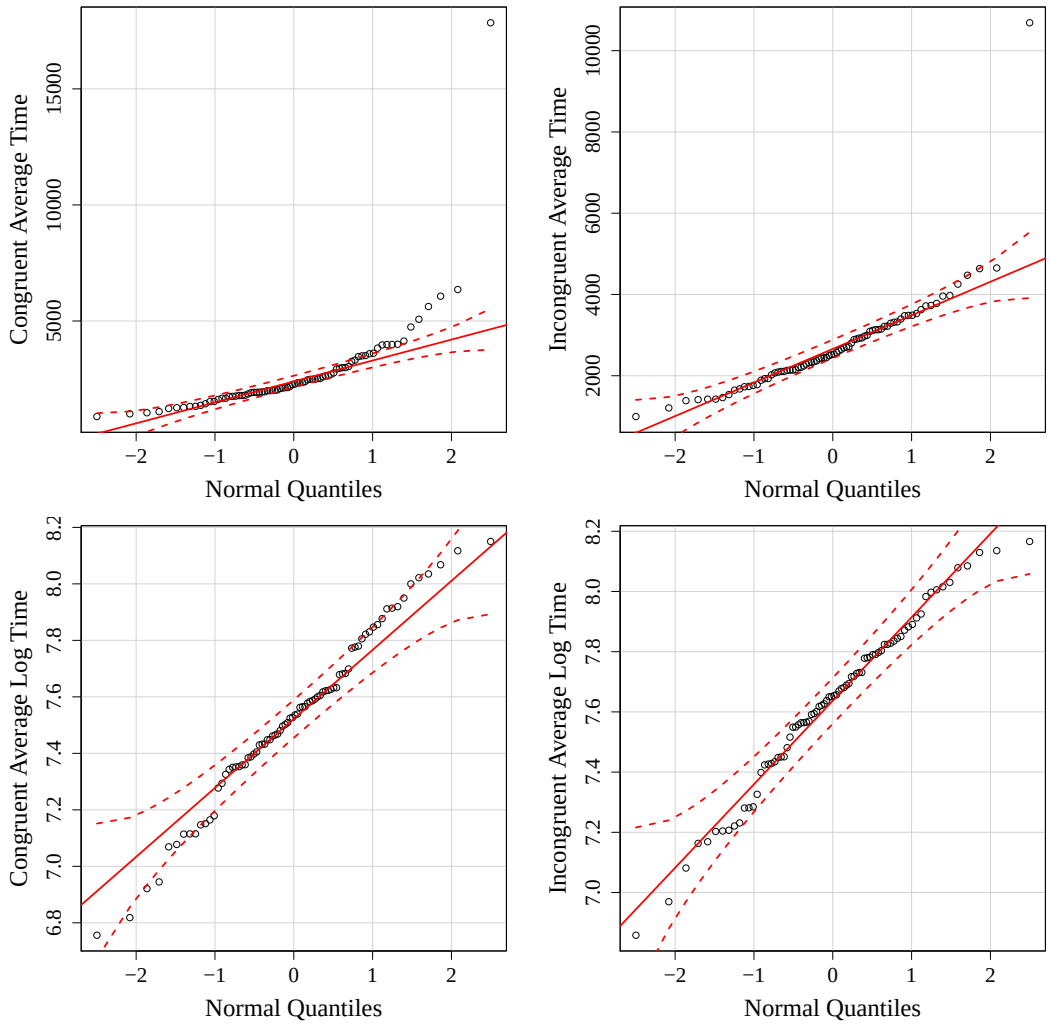


**Figure 4: Normal-QQ plots of the congruent and incongruent reactions times and their logs. The dashed lines indicate 95% confidence intervals about the fitted lines.**

ber of levels solved out of those attempted, and the number of times a subject quit a level without solving it.

Besides the covariates related to the difficulty of the level sets, we recorded many other variables to help account for the differences between players. We narrowed the list down to 28 variables that we felt were likely to have an impact on player performance, including how often they played games, whether or not they were a fan of puzzle games, their subjective impressions of the game, and their scores on the attention test. Principal component analysis was used to further reduce the number of variables to a more manageable level.

There were several other potentially useful covariates that we chose to leave out. Three participants indicated on the survey that they were left handed, and two indicated that they had been diagnosed with some form of attention disorder. Neither of these variables were included due to the low number of data points.

# 3. THE DATA

Before we perform our analysis, a brief discussion of our variables and models is appropriate. This section is divided into three subsections. Section 3.1 describes the independent and dependent variables used in our study. Section 3.2 describes the covariates. Section 3.3 describes the three models for which our analyses will be performed.

## 3.1 Independent and Dependent Variables

There were two independent variables: which level set the participant was playing (either hand made or procedurally generated), and whether that set was the first or second set played. Participants were assigned to the four possible combinations in equal numbers.

There were six dependent variables: congruent reaction time, incongruent reaction time, congruent percentage correct, incongruent percentage correct, congruent percentage unanswered and incongruent percentage unanswered.

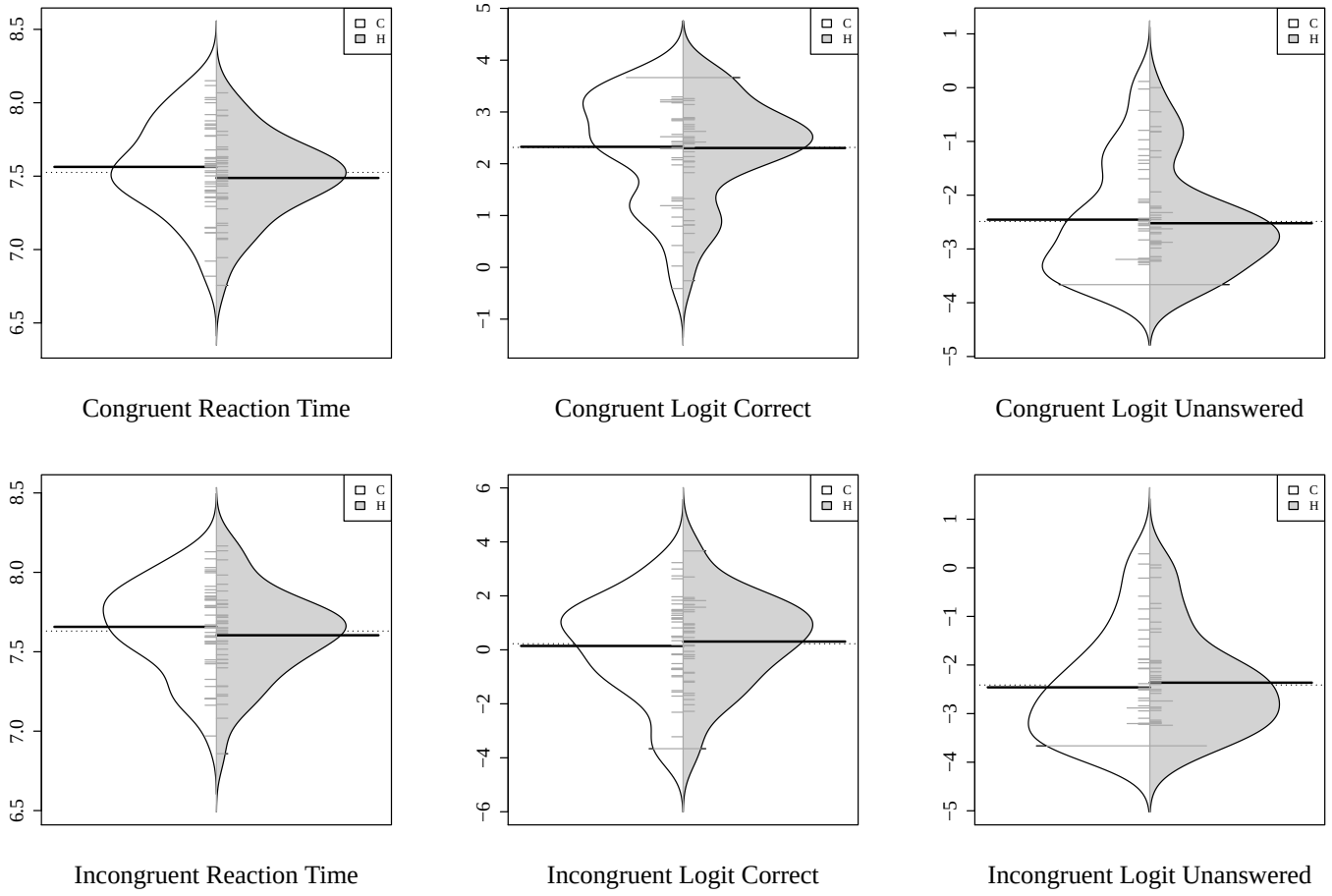


**Figure 5: Bean plots (a form of kernel density plots, see Tukey [21]) comparing the distribution of the scores on the hand made levels and the scores on the computer generated levels. The thick lines show the mean of each subset while the dotted lines show the overall mean for that variable. The shorter gray lines show the data points.**

Reaction times are recorded in milliseconds and log transformed before analysis. On the congruent prompts, the average log response time is 7.53 log-milliseconds (SD = 0.3), or 1.85 seconds. For the incongruent prompts, the time is 7.63 log-milliseconds (SD = 0.28), or 2.06 seconds. Since these are the means of logs, the inverse transformation gives the geometric mean of the original times. (The arithmetic mean of the untransformed times are 2.71 and 2.73 seconds, respectively.)

The percentage measures are transformed using the logit transformation which is in log odds units. For the percentage correct, the log odds are 2.31 (SD = 1.02) and 0.23 (SD = 1.76) for the congruent and incongruent prompts, respectively. This translates to an average of 91% correct for the congruent prompts and 55% correct for the incongruent prompts. For the percentage unanswered, the log odds are -2.49 (SD = 1.02) and -2.41 (SD = 1.07), respectively. This is an average of 7.7% and 8.2% unanswered on the congruent and incongruent prompts, respectively.

These values represent the raw data collected from the experiment before separating the scores on the hand made and procedurally generated levels or adjusting for the covariates. Figure 5 compares the distributions of these six variables and Table 1 summarizes them numerically.

| | | Mean | Std. Dev. |
|---|---|---|---|
| Reaction Time | Con. | 7.53 | 0.30 |
| | Incon. | 7.63 | 0.28 |
| Correct Responses | Con. | 2.31 | 1.02 |
| | Incon. | 0.23 | 1.76 |
| Unanswered Prompts | Con. | -2.49 | 1.02 |
| | Incon. | -2.41 | 1.07 |

**Table 1: A summary of the dependent variables collected for this study.**

## 3.2 Covariates

28 covariates were used in the analysis. These can be logically divided in to 6 groups of variables. Before the primary analysis, we used principal component analysis (see Jolliffe [10]) to reduce the number of these variables. For each group, we kept the principal components with an eigen-

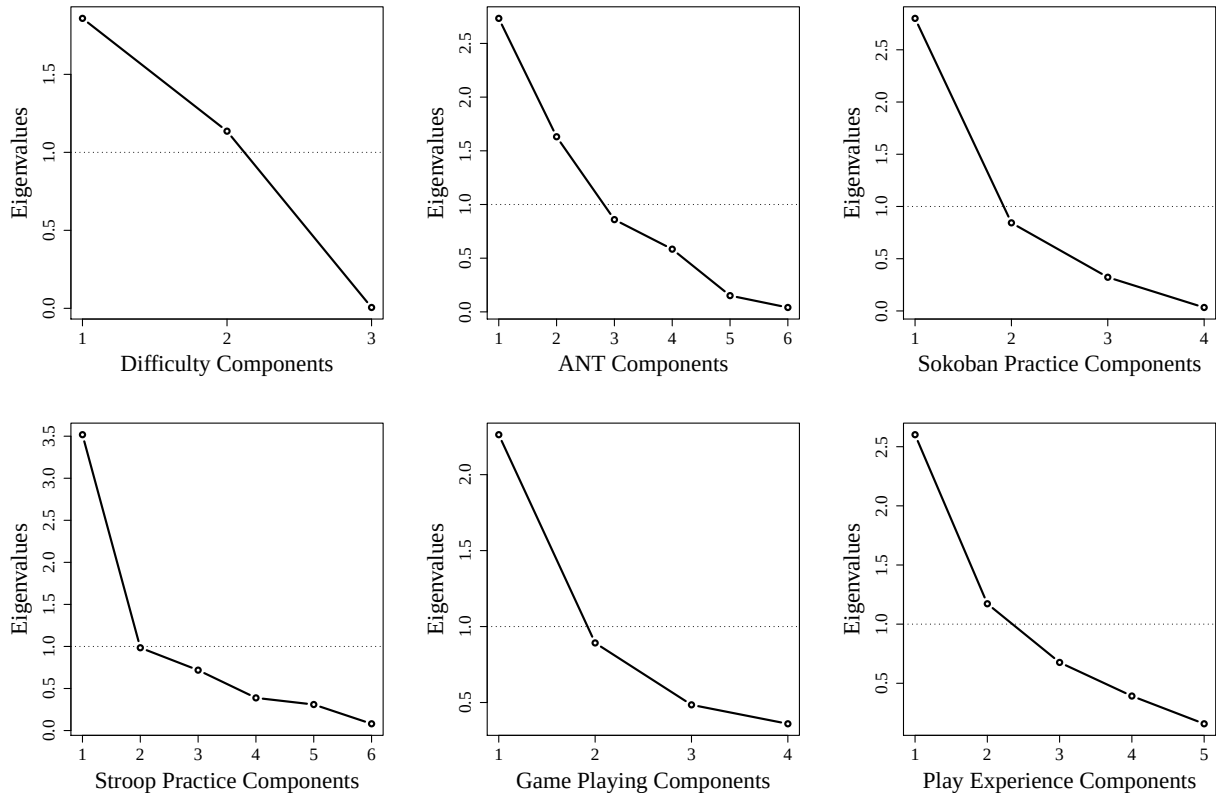


**Figure 6: Scree plots summarizing the number of components extracted from each group of covariates.**

value greater than 1. Figure 6 shows the scree plots for each of the six groups. We used varimax rotation to normalize the components afterwards. We also applied the same analysis to all 28 covariates together.

- There are three variables representing the difficulty of the levels played. These are the number of levels attempted, the number of levels solved and the number of times a player quit a level without solving it. A principal component analysis of these results in two components that explain practically all of the variance of this group of covariates.
- There are six variables representing the participants scores on the Attention Network Test. These are their reaction times and correct response percentages on the congruent prompts, the incongruent prompts and the neutral prompts. (Note that these prompts are not the same as the Stroop prompts used during the game.) The percentages were transformed to log odds before any further analysis. A principal component analysis of these gives two variables that explain 73% of the variance.
- There are four variables taken from the Sokoban practice: the number of practice levels attempted, the number of practice levels solved, the average number of moves beyond the optimal solutions, and the average number of pushes beyond the optimal solutions. The analysis of these variables gives a single component that explains 59% of the variance.
- There are six variables taken from the Stroop practice. These are of the same form as the six dependent variables: the reaction time, percentage of correct answers, and percentage of unanswered prompts, each separated into congruent and incongruent cases. These variables were transformed in the same ways as the dependent variables. One principal component explains 59% of the variance of these covariates.
- There are four variables from the survey that represent the participants' familiarity with gaming and with puzzle games in particular. These are the number of hours per week the participants spend on the computer, the number of hours per week spent gaming, the participants' opinion of gaming on a seven point scale, and whether or not they enjoy puzzle games. The seven point scale used ranged from "Dislike Greatly" to "Neutral" to "Enjoy Greatly." There was a "No Opinion" option, but no one selected it. Whether or not the participants liked puzzle games was taken from a list of game genres they selected their preferences from. One principal component explains 57% of the variance.
- The final group of five variables are from the survey and represent the participants' experiences with the game. Each of these variables are taken from a seven point scale ranging between two antonymous adjectives. These five pairs are "Terrible" to "Wonderful," "Difficult" to "Easy," "Frustrating" to "Satisfying," "Dull" to "Stimulating," and "Boring" to "Fun."

Two principal components explain 75% of the variance within these variables.

We previously mentioned that difficulty does not necessarily correlate with interest, and the analysis of this last group of survey responses provides some evidence for that. One of the two components is largely composed of the "Difficult" to "Easy" value, while the other is largely composed of three of the other four values. ("Frustrating" to "Satisfying" contributes approximately equally to both components.) Additionally, the "Difficult" value and the "Fun" values are almost uncorrelated.

## 3.3 Model Selection

In total, the principal component analysis leaves us with nine variables. The principal component analysis on all 28 variables also leaves us with nine variables (explaining 82% of the variance). From this, we construct three different models and compare the results. All three models are analyzed as linear mixed models with the dependent variables and covariates modeled as fixed effects and the subject ID as a random effect. A compound symmetric covariance structure is assumed. The models differ in the selection of covariates. The three models are:

1. Combine the covariates into logical groups and run a principal component analysis on each group.
2. Run a principal component analysis on all of the covariates simultaneously.
3. Take a single, unrotated component from all the covariates (see Parsons *et al.* [16]).

# 4. ANALYSIS

Tables 2–4 summarize, for the reaction time, the percentage correct, and the percentage unanswered, respectively, the adjusted procedurally generated level scores minus the adjusted hand crafted level scores at a 95% confidence level. Each of these is treated in a separate subsection below.

To get a final result from the three different models, we take the most significant result from each category. This does inflate the significance of the results, and is therefore ill advised for most analyses, but since we wish to show that there is no significant difference between the hand-crafted and the computer-generated levels, inflating the significance only strengthens our conclusion.

The remainder of this section is divided into three subsections. Section 4.1 contains an analysis of the players' reaction times. Section 4.2 contains an analysis of the percentage of correct answers to the auditory Stroop test. Section 4.3 contains an analysis of the percentage of unanswered questions from the auditory Stroop test.

## 4.1 Reaction Time

Reaction time was recorded as the logarithm of the users' times measured in milliseconds. To get an interpretable ratio we invert this transformation by exponentiating the raw differences. Under our assumptions, a slower reaction to the Stroop prompts would imply that the user was paying more attention to that level set. The third model gives the most significant results for the congruent reaction times, and the second model gives the most significant results for the incongruent reaction times.

| Test | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Con. | -0.017 0.177<br>p = .103 | 0.028 0.182<br>p = .008 | *0.023 0.134*<br>*p = .006* |
| Incon. | -0.045 0.149<br>p = .289 | *0.005 0.157*<br>*p = .036* | -0.001 0.111<br>p = .052 |

**Table 2: Reaction time**

For the congruent reaction times, players were on average (95% confidence) between 1.025 and 1.145 times slower in responding to the Stroop test during the procedurally generated levels. For the incongruent times, players were between 1.005 and 1.170 times slower during the procedurally generated levels. This implies that players were paying more attention to the procedurally generated levels and less attention to the Stroop test compared to their times during the hand crafted levels. For comparison, the first model showed no significant difference, while the remaining model showed a less significant result in the same direction.

## 4.2 Percentage Correct

| Test | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Con. | *-0.202 0.700*<br>*p = .272* | -0.305 0.347<br>p = .899 | -0.281 0.264<br>p = .947 |
| Incon. | -0.530 0.424<br>p = .823 | -0.602 0.185<br>p = .294 | *-0.537 0.019*<br>*p = .067* |

**Table 3: Percentage correct**

All percentage data were transformed to log odds before the analysis. To transform to the more interpretable odds ratio, we again exponentiate the raw differences. Under our assumptions, a greater likelihood of getting a correct response to the Stroop test would imply that the players were paying more attention to the Stroop test and less to the game. The first model gives the most significant estimates for the congruent results, while the third model gives the most significant estimates for the incongruent results.

No model gives a significant result at 95% confidence. The most significant of the models for the congruent case shows the players as between 0.817 and 2.014 times as likely to answer the Stroop prompts correctly during the procedurally generated levels (i.e. between 18.3% less likely and 101.4% more likely.) For the incongruent case, the third model gave the most significant result with players between 0.584 and 1.019 times as likely to answer correctly during the procedurally generated levels.

## 4.3 Percentage Unanswered

For the percentage of Stroop prompts left unanswered, a higher score would indicate that the player was paying less attention to the Stroop test and more to the game. The first

| Test | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Con. | *-0.657 0.189* *p = .271* | -0.295 0.323 p = .928 | -0.156 0.353 p = .438 |
| Incon. | *-0.847 0.094* *p = .114* | -0.446 0.243 p = .557 | -0.352 0.218 p = .639 |

**Table 4: Percentage unanswered**

model gave the most significant results for both the congruent and incongruent cases, although none of the results were significant at the 95% level. Players were 0.518–1.208 times as likely to leave a congruent prompt unanswered during the procedurally generated levels, and 0.429–1.099 times as likely for the incongruent prompts.

# 5. CONCLUSION

While their reaction times suggest that the players were paying slightly more attention to the procedurally generated levels than to the hand crafted ones, none of the differences were highly significant. We conclude that players pay about as much attention to procedurally generated levels as they do to hand crafted levels, that they are in a sense *equally engaged*, from which it might be implied that they find both types of level *equally interesting*. Many open questions remain, including the development of other robust measures of the quality of procedurally generated content, and the further elucidation of the relationships between player attention, engagement, and interest.

# References


[1] Douglas G. Altman. *Practical Statistics for Medical Research*. CRC Press, 1990.

[2] Christina M. Armstrong, Greg M. Reger, Joseph Edwards, Albert A. Rizzo, Christopher G. Courtney, and Thomas D. Parsons. Validity of the virtual reality stroop task (vrst) in active duty military. *Journal of Clinical and Experimental Neuropsychology*, 35(2):113–123, 2013.

[3] D. Ashlock and J. Schonfeld. Evolution for automatic assessment of the difficulty of Sokoban boards. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 1–8, 2010.

[4] Joseph Culberson. Sokoban is PSPACE-complete. In *Proceedings of the International Conference on Fun with Algorithms*, pages 65–76, 1998.

[5] Erik D. Demaine. Playing games with algorithms: Algorithmic combinatorial game theory. In *Mathematical Foundations of Computer Science 2001*, pages 18–33. Springer, 2001.

[6] Jonathan Doran and Ian Parberry. Controlled procedural terrain generation using software agents. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(2):111–119, 2010.

[7] Jin Fan, Bruce D. McCandliss, Tobias Sommer, Amir Raz, and Michael I. Posner. Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, 14(3):340–347, 2002.

[8] Petr Jarušek and Radek Pelánek. Difficulty rating of Sokoban puzzle. In *Proceedings of the Fifth Starting AI Researchers' Symposium*, 2010.

[9] Petr Jarušek and Radek Pelánek. Human problem solving: Sokoban case study. Technical Report FIMU–RS–2010–01, Faculty of Informatics, Masaryk University Brno, 2010.

[10] Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

[11] Levent Kirisci and Tse-Chi Hsu. The effect of the multivariate Box-Cox transformation on the power of MANOVA. 1993.

[12] Ruth M. Krebs, Wim Fias, Eric Achten, and Carsten N. Boehler. Picture novelty attenuates semantic interference and modulates concomitant neural activity in the anterior cingulate cortex and the locus coeruleus. *NeuroImage*, 74:179–187, 2013.

[13] Colin M. MacLeod. Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2):163, 1991.

[14] Jeffrey W. MacLeod, Michael A. Lawrence, Meghan M. McConnell, Gail A. Eskes, Raymond M. Klein, and David I. Shore. Appraising the ant: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology*, 24(5):637, 2010.

[15] Jason W. Osborne. Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):1–9, 2010.

[16] Thomas D. Parsons, Albert R. Rizzo, Cheryl van der Zaag, Jocelyn .S McGee, and J. Galen Buckwalter. Gender differences and cognition among older adults. *Aging, Neuropsychology, and Cognition*, 12(1):78–88, 2005.

[17] Wolfgang Schwarz. The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, 33(4):457–469, 2001.

[18] Scott Sinnett, Albert Costa, and Salvador Soto-Faraco. Manipulating inattentional blindness within and across sensory modalities. *The Quarterly Journal of Experimental Psychology*, 59(8):1425–1442, 2006.

[19] J. Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643, 1935.

[20] Joshua Taylor and Ian Parberry. Procedural generation of Sokoban levels. In *Proceedings of the 6th International North American Conference on Intelligent Games and Simulation*, pages 5–12. EUROSIS, 2011.

[21] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[22] Morgan C. Wang and Brad J. Bushman. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3(1):46, 1998.