The Scientific Benefits of Play: Gamifying Sensory Neuroscience

Robert Sekuler¹, Yile Sun¹, and Timothy Hickey²

Abstract
Aim: We present a case study of gamifying research on multisensory (audiovisual) interaction. The advantages and drawbacks of game-based platforms are explored by examining four variants of the same basic experimental protocol.

Method: We implemented and compared four platforms that assessed interactions between auditory and visual processing systems. Two of the variants had been developed in prior research using a minimal to moderate level of gamification to study a particular form of audiovisual interaction. Two new variants were implemented - one with no gamification at all to determine whether subjects preferred the gamified version, and one with much more sophisticated competitive tournament-style gamification to explore the effects of more sophisticated gamification on the experimental design.

Results: All four platforms produced comparable evidence of substantial audiovisual interaction, but subjects preferred the more gamified platforms. Equally important was the efficiency of the most gamified platform: it generated a rich database of ∼40,000 informative stimulus-response events with just three hours’ data collection. This would have taken over 40 hours using a conventional, non-gamified approach.

Conclusions: In platform-to-platform comparison, subjects preferred a gamified platform to one with conventional experimental design. The relaxation of some experimenter control in the most gamified platform had two main effects. It complicated the data analysis because there were more uncontrolled variables to contend with, but exploration of those additional variables opened a valuable window onto a number of unanticipated behaviors and strategies. Importantly, a consistent, strong multisensory interaction was clear even in the most complex, gamified experimental design.

Keywords
perception, gamification, audiovisual interaction, multisensory interaction, cognitive neuroscience

Introduction
Gamification is the process of modifying some activity to include features and principles that are characteristic of games. Gamification can have a variety of positive effects on participants in a gamified activity, including increased engagement and learning (Muthaq et al 2011; Hamari, Koivisto, and Sarsa 2014; Seaborn and Fels 2015; Sailer 2017). However, the effects are highly dependent on the activity being gamified, the characteristics of the participants, and the particular gamification features introduced into the activity. For example, gamification has been used to improve the effectiveness of clinical vision tests for young children who lack the verbal and cognitive skills required in traditional vision tests (Nguyen et al 2014). A comparison of three different versions of a tablet-based color vision test suggested that too much gamification of a clinical screening test might weaken its sensitivity (Bodduluri 2017).

In this paper we present results of a study in which a basic experiment in sensory neuroscience was systematically modified by introducing varying levels of gamification. To generate reliable data, sensory neuroscience experiments must recruit test subjects who will be able to maintain a consistent high level of attention. With no gamification, the repetitive, tedious character of such experiments tends to promote lapses of attention (Abramov et al 1984; Soderquist and Shilling 1992). We were seeking to find evidence that subjects preferred participating in a gamified experiment, even if only a limited amount of gamification was provided. We were also interested in determining whether a fairly large amount of gamification would still allow the fundamental scientific research result to be observed. This research was approved by our institutions Institutional Review Board.

We tested a quartet of related platforms (PF0, PF1, PF2, PF3) for studying audiovisual interactions. The intermediate platforms, PF1 and PF2 employed fairly modest gamification features and had been used to study audiovisual interaction, (Sun et al 2017; Goldberg et al 2015). The platforms differed in their degree of gamification. We expected that increased gamification would motivate

¹Neuroscience Program and Psychology Department, Brandeis University
²Computer Science Department, Brandeis University

Corresponding author:
Timothy Hickey, Computer Science Department, MS018, Brandeis University, Waltham, MA 02453
Email: tjhickey@brandeis.edu
naive, inexperienced research subjects to maintain a high level of attention over the extended testing required by the sensory dimension of our project.

- **PF0** - a conventional laboratory experiment in which a subject would sit in front of a computer screen in a quiet room and make binary category judgments of a sequence of simple visual and auditory stimuli. Responses were signaled by pressing one of two keys on a keyboard. This platform was designed simply to determine whether subjects preferred a minimally gamified platform (PF1) to a non-gamified platform (PF0).
- **PF1** - also a laboratory experiment, but one in which subjects were told they would be playing a game in which they catch “bad” fish and release “good.” As in PF0, responses were signaled by pressing one of two keys. The visual stimuli were cartoon images of a fish, presented against a background of a steadily scrolling image of a river bottom. The images and back-story are the only differences between PF0 and PF1. The results of this experiment are reported in (Sun et al. 2017).
- **PF2** - an experiment carried out in a public space (the Boston Museum of Science) with the game ported to a tablet computer. Rather than pushing a key, a subject tilts the tablet forward to release a “good” fish or backward to catch “bad” fish. As many as two subjects could play at a time and spectators could watch subjects’ progress. Each subject played the game for just five minutes, after which the subject could check a large video monitor showing how their performance compared to that of the previous nine subjects. The results of this experiment are presented in (Goldberg et al. 2015).
- **PF3** - an experiment carried out in an auditorium to which subjects brought their own laptops and headphones or ear buds. Subjects visited a website and played the game using a browser. The game is essentially the same as that in PF1 except that it is structured as a sequence of levels (from 0 to 10) in two different modes (Auditory and Visual). After each game, the subject’s laptop screen displayed a summary leader-board showing the top scorer for each of the ten auditory and visual levels. When the subjects selected a particular level to play, they would also see a leader-board specialized for that level which ranked all of the players for that level. To advance to the next level, a subject had to correctly classify ≥80% of the fish encountered at the current level. At each higher level, both the inter-fish interval and the fish lifetime decreased. Additionally, a summary leader-board was also projected onto a large screen at the front of the room, allowing all subjects to see which competitor was fastest and most accurate at each level and mode of play.

Previously, PF1 was used to study multisensory interaction in young adults (Sun et al. 2017), while PF2 replicated much of the protocol, but with a much wider range of subjects ranging in age from 6 to 82 years (Goldberg et al. 2015). Together, these studies showed that a modest amount of gamification did not obscure the basic audiovisual interaction in which we were interested, and provided evidence that subjects were highly motivated to participate in the gamified experiment.

To probe the limits of this result, we developed two platforms, PF0 and PF3, whose gamification bracketed the modest gamification of PF1 and PF2. PF0 directly tested whether subjects would prefer to take part in a conventional experiment or in a gamified one. Subjects took part in three experiments — first, the conventional experiment, followed by PF1 and finally they were given a choice of PF0 or PF1. As expected, most preferred the gamified experiment. We did not recruit a large number of subjects for PF0 as our goal was to study subject preference (or not) for gamified experiment, not to try to further replicate the well-established results from PF1 and PF2.

Platform PF3 tested whether results from the controlled laboratory version could be replicated more efficiently, in a just few one-hour long game tournaments, and it also explored the effects of adding many more gamification features to the experimental design. As described below, applying exploratory data analysis (Tukey 1977; Behnrens and Yu 2012; Cleveland 1993) to results from the most gamified platform, PF3, revealed important features of subjects’ behavior that would not have been observed with more constrained platforms.

Note that our most gamified platforms incorporated familiar features of video games, such as a story line, interesting visual images, familiar game platforms (tablets and laptops), score keeping and competition. The aim was to increase subjects’ ability to maintain focus and a high level of concentration over a 40 minute session, while maintaining enough control over the experimental conditions to support reliable conclusions about interactions between auditory and visual processing of the stimuli.

**Research Questions**

The research questions for the gamified experiments concern the effects of gamification on the experiment itself. Adding gamification features changes the experimental conditions, thereby likely altering subjects’ behavior. Any data analysis should take account of these possible effects. These changes in experimental conditions also provides a richer data set and exploratory data analysis can be used to make hypotheses about how these gamification features affect the subject’s psychophysical responses.

The main research questions considered in this paper are:

- To what extent do subjects prefer to participate in gamified experiments rather than traditional experiments.
- What are the main advantages and disadvantages of gamifying an experiment in sensory neuroscience?

Our hypothesis was that adding carefully selected gamification features would increase subjects’ engagement and focus, while conserving enough primary features of the experimental conditions to support strong inferences about interactions between visual and auditory processing. We also hypothesized that subjects would prefer to participate in a gamified rather than a conventional experiment.
Design Goals and Decisions

The game versions of the experiment were designed with the following goals:

- The game should be engaging and should encourage subjects to increase their level of concentration during the experiment.
- The user experience for all subjects should be as uniform as possible, even though they might be using different computers.
- The data should provide enough information to produce statistically significant answers to the primary research question about audiovisual interaction.

A secondary goal was to increase the efficiency of the research process both by making it easier to recruit subjects (who may be more inclined to participate in a video game experiment than an audiovisual perception experiment) and by allowing the experimenter to supervise multiple subjects at the same time.

Successful video games incorporate a number of essential components (Straat, Rutz, and Johansson 2014). Table 1 summarizes, for some of those essential components, the difference between the most gamified of our platforms, PF3, and a non-gamified platform, PF0.

PF0: the conventional design

The conventional version of the experiment, PF0, had a highly regulated environment with a very limited range of responses the subject could make to a similarly limited range of stimuli.

The subject sat in a room viewing a computer display. The experimenter explained that the subject would be presented with a sequence of images and sounds. The images comprised a red disk whose size oscillates at either a slow rate (5 Hz) or a faster rate (8 Hz). The disc remained centered on the display. Fig. 1 shows a snapshot of the stimulus. The images appeared on a solid black background and there was a continuous sound (similar to water flowing in a stream) as a background sound.

A synthesized, multifrequency, broadband sound was presented concurrently with the oscillating disc. The sound oscillated in amplitude at either a slow rate (5 Hz) or at a faster rate (8 Hz). The experiment had two phases: one, the Visual mode, in which subjects were instructed to base their judgments solely on the rate of stimulus oscillation, for such stimuli we hypothesized that the response time would be shorter and the accuracy higher than when they were unisensory conditions by pressing the “O”-key without regard to the rate of stimulus oscillation, for such stimuli we hypothesized that the response time would be shorter and accuracy would be higher.

Analysis of the data

The experiment included two modes: Visual and Auditory, where the names refer to the critical sensory stream, visual or auditory, on which subjects were instructed to base their judgments. Table 2 summarizes the conditions. Each mode, Visual or Auditory, includes eight conditions, comprising four combinations of fast or slow visual modulation (Fv or Sv), paired with fast or slow auditory modulation (Fa or Sa), plus four unsensory control conditions. We further designate a fish an Oddball if it is visible but mute, or if it is invisible but audible. We classify a fish as either Congruent, when its auditory and visual modulations are at the same rate, or Incongruent, when its auditory and visual modulations differ in rate.

For each mode, Auditory or Visual, the events over all subjects were then filtered by the stimulus type and the mean response time and accuracy were calculated for all such events. The hypothesis was that when the visual and auditory events were congruent (i.e. FvFv, or SvSv), response time would be faster and the accuracy higher than when they were incongruent (i.e. FvSv, or SvFv). As subjects were to respond to unsensory conditions by pressing the “O”-key without regard to the rate of stimulus oscillation, for such stimuli we hypothesized that the response time would be shorter and accuracy would be higher.

Advantages of the Conventional Platform

This experiment was relatively easy to create using Matlab and the task was easily understood by subjects. All subjects saw and heard the same stimuli as they were seated in front of the same computer in a quiet room. There were eight different conditions which were presented in random order. The subjects made about 20 classifications per minute over the course 20 minutes for each of the two modes. This produced ~400 events for each mode and each subject.

Disadvantages of the Conventional Platform

Each experiment with the conventional platform required one hour of the experimenters’ time to collect about 800 classifications per subject. As a result, if an experimental design demanded 40,000 classifications, 50 hours of experimenter time would have been required.

Also, the experiment requires multiple repetitions of the same conditions, which opens the possibility that subjects will become bored and lose focus, especially toward the experiment’s end. This can pose a challenge for many conventional, non-gamified experiments in behavioral neuroscience.
Table 1. Comparing the Four Platforms

<table>
<thead>
<tr>
<th>Component</th>
<th>Gamified Platform PF3</th>
<th>Conventional Platform PF0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story</td>
<td>Narrative Backstory</td>
<td>Task-focused</td>
</tr>
<tr>
<td>Rules</td>
<td>Track Subjects’ success</td>
<td>Static</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>Realistic</td>
<td>Abstract</td>
</tr>
<tr>
<td>Technology</td>
<td>Subject’s own device</td>
<td>Experimenter’s device</td>
</tr>
</tbody>
</table>

Table 2. Summary of Conditions

<table>
<thead>
<tr>
<th>Type</th>
<th>Label</th>
<th>Designation</th>
<th>Visual (Hz)</th>
<th>Auditory (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bisensory Conditions</td>
<td>Congruent</td>
<td>$F_v, F_a$</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Congruent</td>
<td>$S_v, S_a$</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Incongruent</td>
<td>$F_v, S_a$</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Incongruent</td>
<td>$S_v, F_a$</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Unisensory Conditions</td>
<td>Oddball</td>
<td>$F_v$</td>
<td>8</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Oddball</td>
<td>$S_v$</td>
<td>5</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Oddball</td>
<td>$F_a$</td>
<td>–</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Oddball</td>
<td>$S_a$</td>
<td>–</td>
<td>5</td>
</tr>
</tbody>
</table>

PF1: A Minimally Gamified Experimental Platform

In the first gamified version of the experiment, the only changes to the conventional design were the introduction of back-story to motivate their behavior and the replacement of the simple geometric stimulus on a solid background with an image of a fish on a scrolling background of a mossy river bottom. We begin this section with a review of the methods and results from this experiment, as reported in (Sun et al 2017), and then discuss the advantages and disadvantages of gamification in this experiment.

For this game we created a simple but engaging backstory, which we explained to subjects at the beginning of the experiment. This backstory was designed to provide motivation for their actions, and memorable, easily understood rules of the game. These rules included the basis on which judgments were to be made, and the keyboard keys that were to be pressed in order to signal those judgments. The subjects were told that they would serve as fish police, charged with cleaning "bad", invasive fish out of a river, while allowing "good", native fish to remain. They would classify each fish by pressing one of three keys: U, I, O. Whenever they saw a bad fish, they were to press I (for Invasive species); whenever they saw a good fish, they were to press U (for Usual species). An Oddball fish, one that didn’t fit either category, was classified using the O key. The fish would oscillate visually by having their vertical scale increase and decrease sinusoidally at either 6 or 8 Hz. While a fish was on the screen, the subjects heard an amplitude modulated sound oscillating at either 6 or 8 Hz. Subjects played the game in two modes: Auditory or Visual. In Visual mode, the "bad," invasive species were the ones whose images oscillated rapidly (8 Hz). In the Auditory mode, the "bad" fish were the ones with the rapidly oscillating sound. In either mode, a silent fish or an invisible fish was an oddball and classified using the O key.

Figure 1 shows a typical screen from the game. The fish were released from either the left or right side of the screen.
and swam across the screen at a steady rate. After two seconds they disappeared, making way for a new fish to appear after an inter-fish interval of 1 to 2 seconds. When a subject correctly classified a fish, a distinctive "success" sound was played; incorrect classifications triggered a different, "failure" sound.

**Method**

Twelve subjects were recruited and each participated in a 30 minute session in a quiet room. During the first 10 minutes, the experiment was explained and informed consent was obtained. This was followed by about 20 minutes of responding to stimuli presented on a desktop computer. Each of the 12 subjects was presented with 480 stimuli spread over six blocks of trials. These blocks included two each of congruent, incongruent, and control conditions. Each of the two congruent blocks consisted of 80 stimuli that consisted of \( F_v F_a \) or \( S_v S_a \) in random order. The incongruent blocks had 80 \( S_v F_a \) or \( F_v S_a \) again with 40 each of these two types. The control blocks had 40 \( F_v \) and 40 \( S_v \).

As the hypothesis was that subjects would be faster and more accurate on congruent as opposed to incongruent stimuli, the accuracy and response time for stimuli was calculated separately for each of the three types of blocks - congruent, incongruent, and control.

**Analysis of the data**

First, mean accuracy for congruent fish was 0.86, compared to 0.71 for incongruent fish, a statistically significant difference \((p < 0.001)\). Moreover, mean response time for congruent fish was 853 ms, compared to 945 ms for incongruent fish. This result, too, was statistically significant, \(p < 0.02\). Accuracy and response time for the control fish were essentially indistinguishable from those for incongruent fish, but they differed significantly from results with congruent fish, \(p < 0.02\).

**Advantages of PF1’s Gamification**

There were two main advantages of this approach. The first is that the experimental design maintained a high level of control over the experience of the subjects. They were each exposed to the six different kinds of stimuli in a set order while sitting in a quiet room with no other distractions. The second advantage is that the data analysis was very straightforward. Each subject responded to 160 stimuli of each of the three types - Congruent, Incongruent and Control. This allowed us to demonstrate that in auditory and in visual mode, there was a statistically significant difference between the congruent and incongruent responses both for accuracy and for response time, as was hypothesized. These results as well as additional analysis were published in (Sun et al 2017).

A few of the subjects asked if they could continue playing the game after the experiment was over, indicating that they found the experience rewarding in some manner.

**Disadvantages of the PF1 Gamification**

PF1’s main disadvantage was that it failed to engage subjects as fully as we would like. Specifically, subjects sat in a quiet room, observed and categorized 480 fish presented one after another, one appearing about every 3 seconds.

Task difficulty, as indexed by the difference between a fish’s two possible oscillation rates, was held constant. Additionally, PF1 lacked a social dimension that could have increased engagement: there was no leader-board and no feedback about how the subject’s performance compared to that of other subjects. The gamification was modest: PF1 was accompanied by a backstory in which the subject would police the river, and included the scrolling stream-bed background and cartoon fish images that oscillated and moved across the screen.

**Comparing PF0 and PF1**

We tested the hypothesis that subjects would prefer a gamified version over the traditional platform, even if the level of gamification was modest. For this purpose, we recruited six subjects for a session in which each first played PF0 and then PF1, or vice versa. Then they played a third game for which they were free to choose to play either PF0 or PF1. Note that our goal was not to replicate the detailed results of PF1, but to test our hypothesis that subjects would prefer PF1 to PF0.

The outcome of this head-to-head comparison between platforms was clear: Every subject chose to play the more gamified version (PF1), showing that subjects did indeed prefer the more gamified version, even if its gamification was modest.

**PF2: The Museum of Science protocol**

For PF2 we ported the FishPolice!! game to an Android tablet as shown in Fig. 3. Subjects would hold the tablet with both hands, tilting it backward, toward themselves, for bad fish (as though to "catch" the fish), or tilting the tablet forward for a good fish (as though to release the fish). This experiment was carried out in the Boston Museum of Science’s Living Laboratory facility. A large display screen above the experiment’s site showed an image from the game. The very limited time made available by the Museum staff meant that data collection, which could be done only for 90 minutes once a week, had to be spread out over several months. During that time, 60 subjects, ranging in age from 6 to 82 participated. The subjects first supplied some demographic information and gave written informed consent. They were then given a short training session, which was repeated if they did not attain sufficient accuracy during the training. After training each subject played the game for about five minutes, during which 60 successive fish were classified as good or bad. Note that the platform was essentially the same as in PF1 except that the game had been ported to a tablet which necessitated a different response mechanism; rather than pressing a key on a keyboard, subjects tilted the tablet in one of two directions, either toward or away from themselves. Also, the game included no oddball/control fish, only congruent or incongruent ones. Additionally, only a Visual mode was tested, with subjects judging whether the fish oscillated in size slowly (good fish) or rapidly (bad fish).
Analysis of the data
The data analysis was essentially the same as in PF1. Subjects were able to classify congruent fish significantly more accurately and quickly than incongruent fish. Details of the results of the experiment and a detailed analysis can be found in (Goldberg et al 2015).

Advantages of the PF2 Gamification
This experiment validated our original findings for a sample with a much wider age range with a smaller time requirement for each subject (10 minutes as opposed to 30 minutes). This platform was significantly more engaging than PF0 or PF1. Evidence for this greater level of engagement is that although subjects were not being compensated for their time, only about 10% quit before completing the game, and about half of those left because of extenuating circumstances (e.g. a parent insisting that their child had to stop playing the game.) Moreover, about a quarter of the subjects asked if they could play again. Those requests were almost always denied because there was long line of volunteers waiting their turn.

Also, data could be collected for two subjects at a time as we had two tablets available. There was a leader-board that showed the scores of the five most recent subjects, providing a modicum of competitive stimulus.

Disadvantages of the PF2 Gamification
The main disadvantages were the loss of control of the environment. The experimental venue was noisy, with many visual and social distractions. Moreover, because a subject held the tablet during testing, the viewing distance and light reflected from the tablet differed somewhat among subjects. In spite of these variables, though, the hypothesized experimental effect was still very strong. Another disadvantage was that the subjects were unpaid museum visitors and they were only able to interact with the game for five minutes each, just a quarter of the time each subject played the game in PF1. This required four times as many subjects to get an equivalent number of stimulus-response interactions, and the reduced time available for each subject meant a larger percentage of the available time had to be spent on obtaining informed consent and acquainting the subject with the task. Because of the museum-imposed limits on experiment time per week, it took several months to generate enough data to observe statistically significant effects.

PF3: The most gamified tournament-style, browser-based platform.
The most gamified experiment using PF3 exposed the subjects to an environment that differs in several ways from the less gamified platforms. This section discusses the similarities and differences between the most gamified platform (PF3) and the three more traditional platforms (PF0, PF1, PF2).

Visual and Auditory Stimuli
The visual and auditory stimuli in PF3 are exactly the same as in PF1 and PF2.

Levels and Timing
One of the biggest differences between the PF3 and the other platforms was its introduction of different levels of difficulty. Subjects started at the training level (level 0) in which there was a larger difference between slow and fast oscillations –3 and 8 Hz rather than the 5 and 8 Hz oscillations used in subsequent levels. This larger difference made the two rates more discriminable, thereby easing subjects’ entry experience. In levels 0-3 there were no oddball fish, only congruent or incongruent. In levels 4-10, fish could be either congruent, incongruent, or oddball. Oddball fish were generated with a probability of 20%, while congruent and incongruent fish appeared with probability 40% each. The faster and slower versions of each stimulus were equally probable.

Fig. 4 shows the leaderboard for PF3. Subjects saw this leaderboard on their laptop at the start of the game, and after every game thereafter. The leader-board was also displayed on a large screen at the front of the auditorium where game tournaments were held for PF3. Initially, subjects could only select level 0 of the Auditory or Visual mode. Each time they achieved 80% accuracy or more on a level, the next level became available.

The leader-board served two key psychological purposes. It encouraged competition, and it provided subjects with a visual indicator of their progress in the experiment, showing clearly which levels they had mastered in each mode, Visual or Auditory.

The most important ways in which the levels differed were in the amount of time subjects had to classify the stimuli (hereafter, fish lifetime) and in the average time between a subject’s response and the appearance of the next stimuli (hereafter, inter-fish interval).

In PF3, fish appeared at a uniformly random time within a specified interval after the previous fish has disappeared. This inter-fish interval (IFI) decreased systematically over successive levels of the game, which provided an increasing challenge to the subject. The interval between the disappearance of a fish and the spawning of its successor was selected uniformly randomly from the interval $I_L$ which was given by the following formula in terms of the level $L$, which varies from 0 to 10:

$$I_L = [1000 - 75 \times L, 2000 - 150 \times L]$$
With this formula, the interval $I_L$ started at 1000 to 2000 ms for level 0, and decreased to 250 to 500 ms, for level 10.

The fish disappeared when the subject made a classification response by hitting one of the U, I, or O keys. If the subject failed to hit any key before the fish lifetime expired, the fish also disappeared. The fish lifetime depended on the level $L$, which started out at 2000 ms, and dropped 150 ms for each successive level until it reached 500 ms for level 10.

**The GamePlay Plot**

The introduction of levels and subject agency into the game complicated the analysis of the data. Subjects all started at level 0 and could advance to the next level only if they had achieved at least 80% accuracy on the previous level. The game allowed them, however, the flexibility to stay on the current level even if they were qualified to move to the next; they could also drop down to a previous level. Additionally, they had the option of either spending time looking at the leader-boards, or proceeding immediately to the level.

We expected most subjects to simply advance up the levels in visual or auditory mode until they reached their limits, which we expected to vary between subjects. To test this expectation and to track individual subjects’ trajectory through game space, we generated GamePlay plots, a visualization inspired by the methods of exploratory data analysis (Cleveland 1993). We devised this graphical device believing that, as Tukey wrote, "The greatest value of a picture is when it forces us to notice what we never expected to see."

Fig. 5 is a zoomable GamePlay plot generated using a Matlab script that shows the progress of Subject 9 through successive levels of FishPolice!!, with performance in Visual mode in blue and performance in Auditory mode in red. The horizontal axis presents the elapsed time in minutes from the start of the game, and the vertical axis represents the level.

By zooming into the plot, additional fine structure is revealed. For example, Fig. 6 shows a zoomed in view of the first five minutes of Subject 9’s game play. The zoomed in area is shown as the rectangle in the lower left corner of Fig. 5, and Fig. 7 shows a further zoom into the first full game. At the zoomed in levels the vertical axis encodes both the level and the response time.
The horizontal axis of the GamePlay plot shows the number of minutes since the experiment’s start. The vertical axis is more complex and encodes both the level and the response time. In the fully zoomed out mode, it encodes the level of each game. In the fully zoomed in mode, it encodes the reaction time for each individual response. In the intermediate zoom modes, it displays both the level and the reaction time simultaneously. Data points represent correct responses in Visual mode (blue diamond) or Auditory mode (red diamond) and incorrect responses in either Visual or Auditory mode are (magenta asterisks, centered at (x,y)) where x is the time of the subject’s response, and y represents level and response time, as described above. In a subject’s full game play plot it is difficult to see the incorrect responses without zooming in, as the incorrect responses are obscured by the numerous correct responses.

The full GamePlay plot for a subject provides a detailed overview of their overall progress in the experiment. For example, from Fig. 5 shows that Subject 9 advances to level 9 in Visual mode in about 11 minutes, but then is stuck on that level for the next eight games. After the first half of the experiment ends at minute 20, the subject takes a few minutes to switch over the Auditory mode, but then steadily advances through that mode during minutes 22 to 33, finally reaching level 8 after having had to repeat only one level (level 2). Subject 9 then stays at level 8 for next seven minutes (and seven games).

The higher levels require the subjects to classify more fish, more rapidly, with less time between each classification which is why we see subject 9 plateauing at levels 9 and 8 in visual and auditory mode, respectively. At level 0 the subject has two seconds to make a decision, while at level 9 the response must be made before 650 ms have elapsed. Similarly, at level 0 the inter-fish interval averages about 1.5 seconds, while in level 9 it is close to 0.5 seconds (487.5 ms) which causes many more fish to be presented during a one-minute game at level 9. In particular, level 0 play can entail as few as 18 fish per game, while ~50 fish are presented for classification during each one-minute game at level 9. At the higher levels we expected that these constraints would require the subjects to be highly engaged and focused so they could perform at their highest level.

Fig. 6’s zoomed-in view for subject 9 shows that most of the subject’s responses were in the range 0.4-0.8 seconds as the minor grid lines on the vertical represent 0.4 second intervals. The subject had up to two seconds to respond, so all responses will be between the major grid lines as they represent 5*0.4=2.0 seconds of possible response time. Also, subject 9 made four incorrect categorizations but this visualization makes them hard to discern.

The higher level of zoom in Fig. 7 clearly reveals the 29 individual classifications and their response times for the first game (at level 0) of subject 9. In particular, we see that the response times were all between 400 and 800 ms and 3 of the 4 mistakes were near the 400 ms limit while the last was closer to 750 ms. The subject correctly identified 25/29 = 86.2% of the stimuli and so was able to move on to the next level in game 2.

Game Playing Strategies

When we analyzed the GamePlay plots for all 43 subjects, we discovered that the systematic, steady progression from level to level that Fig. 5 showed for subject 9 was the exception rather than the rule. Fig. 8 shows the GamePlay plots for subjects who were tested first in the Auditory mode; Fig. 9 shows the GamePlay plots for subjects who were tested first in the Visual mode. The plots reveal a wide variety of skills and tactical behaviors among the 43 subjects.

Several subjects succeeded in almost every game they played, advancing level by level until they reached a level beyond which they could not advance, even after many tries. Subject 9 is one example of this steady progress. Most subjects, though, exhibited more complex, sometimes unexpected, behavior.

For example, we can see, in Fig 10 that subject 22 had difficulty in Auditory mode. It took three attempts at level zero before reaching the 80% accuracy level required to advance. Subject 22 played 3 games at level 1 without being able to advance and then chose to drop back down to level 0 for game 7. In games 8 through 14, subject 22 advanced to level 4 in Auditory mode. At this point, the subjects were asked to switch modes. Subject 22 seemed to find Visual mode much easier, advancing from levels 0 to 7, and meeting the criterion of 80% accuracy on the first attempt for nearly all of those levels. However, level 7 seemed to represent a ceiling for that subject, who remained at level 7 for games 10-15 in Visual mode.

Zooming in on individual games in a GamePlay plot, reveals the fine structure of play, including the details of timing, whether individual responses were correct or wrong, and the response times associated with those individual responses. For example, Fig. 11 shows subject 22’s first four games. This figure uses the same plotting conventions as in Fig. 10, except that at this finer scale the timing of each stimulus and response is visible. Successful classifications of fish are indicated by diamonds (red or blue depending on the mode) and incorrect classifications are magenta asterisks. We can see that over the 25 fish in game 3, the subject responded made only four incorrect responses, for an accuracy of 84% (21/25). This allowed the subject to advance to the next level for game 4. We can also see that the incorrect responses occur for a range of response times and include responses with the shortest as well as ones with the longest response times. Incorrect responses are also spread throughout the game and do not appear clustered near the beginning or the end. Finally, we see that the number of incorrect responses decreased from game 1 to game 3, but when advancing to level 1, accuracy dropped back to ~50-60%. The main difference between level 0 (a training round) and level 1 was that a fish’s faster and slower oscillations were easier to distinguish at level 0 compared to all higher levels. Level 0’s rates were 3 and 8 Hz, compared to 5 and 8 Hz at higher levels.

Other Strategies

Subject 13 (row 2, col 2 of Fig. 8) shows a behavior similar to Subject 22, but only reaches level 3 in visual and auditory. This subject needed ~6 or 7 games to reach their maximum level. After that, the subject could not advance, and played the next few games (6-8) at level 3. This subject
Figure 8. Visualization of GamePlay for subjects who were tested first in the Auditory mode (shown by red markers), and then in Visual mode (shown by blue markers). The Horizontal axis is the time, in minutes, since the beginning of the experiment. The vertical axis represents the level, with small perturbations from the response time of each individual event. Note that this detailed, event-by-event information is not visible unless one zooms into the figure, as is done below, for Figs. 9 and 11.

Figure 9. GamePlay Visualizations for subjects who were tested first in the Visual mode and then in the Auditory mode. The plotting conventions are the same as in Fig. 8.
also demonstrates an interesting strategy, by dropping back to level 0 in game 11 of the Auditory mode. This retreat might have been done to build confidence after failing several attempts to advance beyond level 3, could have an attempt refresh their memory of the difference between oscillation rates.

Subject 32 (row 5 col 2 of Fig. 9) reached a maximum of level four in both Visual and Auditory modes, but needed two to four games at each level in order to advance. This suggests that as both the inter-fish interval and the time allowed for response decreased, the subject required a few games at each new level to learn how to cope with the increased time pressure.

Next, observe that some subjects showed an asymmetry between their performance in Auditory and in Visual modes. For example, subject 15 (row 2, col 4, Fig. 9) reached level 8 in Auditory mode but could only reach level 5 in Visual mode. This difference could have been caused by lack of time, or by a difference in the subject’s ability to discriminate auditory vs. visual rate.

There is also some evidence that some subjects had some initial difficulty in auditory mode. In particular, subjects 2 (row 1 col 2 of Fig. 9) and 12 (row 2 col 3 of Fig. 9 took several attempts at level 0 in auditory mode before finally advancing rapidly). A more extreme case is Subject 25 (row 4 col 2 of Fig. 9). That subject reached level 7 relatively quickly in Visual mode, but did not advance past level 0 in Auditory mode. It could be that the subject misunderstood the rules in Auditory mode, or it could be that the subject was simply unable to distinguish between fast and slow auditory oscillations even at the initial, 3 and 8 Hz rate difference presented at level 0.

We see one example of a subject in visual mode having some initial difficulty with Subject 6 (row 1 col 3 of Fig. 8 who started off in visual mode then switched to auditory mode, and after 20 minutes went back to visual mode but was unable to advance past level 1 for several games. The subject then took a few minutes break and was able to advance to level 3.

Subject 41 (row 5 col 4 of Fig. 8 shows an interesting behavior: in Visual mode, the player interrupts their progression to higher levels to drop down to retry previous levels on which they already had had success. Our working hypothesis is that this player was attempting to be the highest scorer at those previous levels, even though this would make it less likely they would become the top scorer overall (which carried a $25 cash prize). We see this same phenomenon in several other subjects, e.g. subjects 5, 13, 33 in Fig. 8 and subjects 7, 37 in Fig. 9.

Subject 43 (row 2 col 4) exhibits a variation of this behavior where they interrupt their progress in Auditory mode to return to their top level in Visual mode, possibly hoping to break the record and become the top scorer. This was unexpected behavior, but it does make sense given the incentives of the experimental design and subjects’ agency to choose the mode and level of their games. Subject 1 in Fig. 8 exhibits the same behavior.

Finally subject 1 played a few games in Auditory mode, and then switched to Visual mode. It is not clear why they made this choice, but it could be that they found the Auditory mode too difficult and chose to try Visual mode instead. Subject 6 exhibited the same behavior, but in the opposite direction, starting with Visual mode and then switching into Auditory mode.

There are many other observations we could make about the behavior of subjects in terms of their agency, but the main lesson about adding agency into a gamified version of an experiment is that there is a strong chance that subjects will use their agency to engage in behavior that the experimenter did not anticipate. In our case, subjects who used their agency in different ways nevertheless made approximately the same number of attempts in both Visual and Auditory modes, but their trajectories through those modes and levels were non-monotonic. Sometimes, subjects dropped down to a previous level or returned to the previous modes.

We could have curtailed this variability by requiring subjects to play at the highest level for which they had qualified, keeping them from shifting to the other mode (Auditory from Visual, or vice versa) until they had played...
Table 3. Accuracy from Fully Gamified Platform (PF3): Visual Mode

<table>
<thead>
<tr>
<th>Condition</th>
<th>Accuracy</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congruent</td>
<td>85.5%</td>
<td>[84.7%, 86.2%]</td>
</tr>
<tr>
<td>Incongruent</td>
<td>80.5%</td>
<td>[79.6%, 81.3%]</td>
</tr>
<tr>
<td>Oddball</td>
<td>82.0%</td>
<td>[80.7%, 83.2%]</td>
</tr>
</tbody>
</table>

Table 4. Response Times from Fully Gamified Platform (PF3): Visual Mode

<table>
<thead>
<tr>
<th>Condition</th>
<th>Accuracy</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congruent</td>
<td>731.4 ms</td>
<td>[727.8 ms, 734.9 ms]</td>
</tr>
<tr>
<td>Incongruent</td>
<td>747.2 ms</td>
<td>[743.2 ms, 751.1 ms]</td>
</tr>
<tr>
<td>Oddball</td>
<td>679.5 ms</td>
<td>[674.1 ms, 685.0 ms]</td>
</tr>
</tbody>
</table>

...for 20 minutes. However, the actual game play plots suggest that subjects chose to drop to lower levels in an effort to improve their performance, which served the main goal of our gamification, namely, allowing subjects to operate at what they considered their own optimal level.

Results

Our first observation is that the PF3 experiment showed a similar effect of congruence on the accuracy and response times of subject responses that we had observed in PF1 and PF2. This can not be considered a replication of the original results though because the subjects in PF3 were subjected to time pressure as they increased from level 0 toward level 10. The higher levels required them to make responses more quickly and required them to classify many more stimuli as the inter-fish interval decreased. The shorter inter-fish intervals both gave the subjects less time to prepare for the next stimulus and required them to classify more stimuli in the one minute each game comprised.

Aggregating subjects’ responses across all levels beyond the training level (i.e., for levels 1-10), yields the results shown in Tables 3 and 4. These aggregated results confirm the basic result from the other platforms – players are more accurate and fastest when classifying congruent fish, and less accurate and slowest when classifying incongruent fish.

Advantages of PF3 gamification

This platform confers several advantages.

- Efficiency Because of the tournament style venue, PF3 allowed a large amount of data to be gathered rapidly by having a potentially large number of subjects participating simultaneously. In three one-hour sessions, ~40,000 stimulus-response events were collected from 43 subjects. Gathering this amount of data would have required a full week of 10 hour days using the less gamified Platforms.

- Additional Variables Introducing multiple levels of difficulty had the effect of expanding the scope of the experiment to also include an additional variable – achievement level. The level combined both time pressure and additional practice. Each subject was exposed to up to 11 different variants of the experimental conditions (depending on the level). All subjects started at level 0 and could only advance to the next level if they achieved an accuracy of at least 80%. By restricting analysis to all subjects at a single level, the experiment allowed the effect of time pressure in a training model of performance to be studied. The analysis was complicated by the fact that subjects have to work their way up to higher levels of time pressure. As a result, there was no easy way to separate the effects of learning from the effects of time pressure. Moreover, time pressure was induced by two processes – decreasing the time allowed for a subject to respond, and also decreasing the inter-fish interval between a response to one fish and the appearance of the next fish.

- Less Time Variance As all of the tournaments took place on the same week day at the same time of day, potential confounding effects such as time of day or week were removed. Achieving this by testing one subject at a time would take weeks for a single researcher to collect this quantity of data.

Disadvantages of PF3 gamification

The introduction of competition and levels into the game had several effects that complicate data analysis.

- Competition The effects of competition are difficult to judge. Although some subjects self-reported high levels of engagement while they were trying to win the tournament, it is possible that other subjects may have felt discouraged at seeing others performing at a much higher level.

- Agency Although PF3 allowed the subjects to choose to return to an easier level, which could have complicated the analysis, this was a relatively rare occurrence. Among the 43 subjects, at some point in the experiment, 11 chose to regress, retreating to an easier level. We can only speculate on the reason for this regression. Sometimes, the decision seemed to be motivational: a strategy to reach the top of the leader-board for some previous level. Other times, this maneuver may have more task-related: unable to advance a level, subjects may have regressed to refresh their memory of the two different modulation rates.

- Equipment differences Allowing subjects to use their own computers and headphones/ear buds made the experimental conditions somewhat different for each subject; of course this carried a countervailing advantage – it allowed subjects to use equipment with which they were familiar and comfortable.

- Additional Variables The analysis is complicated by the fact that subjects had to earn their way up to higher levels of time pressure. As a result, there was no easy way to separate the effects of learning from the effects of time pressure. Moreover, the time pressure was controlled by two perfectly correlated variables. From one level to the next, fish lifetime and inter-fish interval decreased in lockstep. As a result, our design did not afford a way to separate out the effects of these two processes; that would require breaking the correlation between the two variables in order to determine the relative importance of each.
• Multiple-levels Having the subjects move from level to level changed the experimental conditions across games, making it somewhat challenging to compare results across games. Different subjects advanced to each level after different numbers of games and hence had more (or less) training before working on that level. Moreover, having many possible levels decreases the amount of data collected for any one of experimental conditions, which made it more challenging to get results that were statistically reliable for any particular level.

Conclusion

The experiments described in this paper demonstrate that gamification can be an effective tool in sensory neuroscience, and that a researcher has a wide variety of choices when incorporating gamification into a traditional experimental design. The one caveat to this statement is the likelihood that this approach would only work when applied to a robust phenomenon. If the target perceptual or cognitive effect were weak or subtle, that effect might be overwhelmed unless studied in an experimental setting whose conditions were tightly controlled and in which potential confounding variables, such as variations in extraneous environmental stimuli, were minimized. Of course, one might argue that a phenomenon detectable only under tightly controlled conditions is likely to be inconsequential when subjects operate in their typical daily environments.

The level of gamification in the platforms PF1, PF2, and PF3 was relatively modest. PF1 merely added a backstory which was supported by a scrolling background and a cartoon image of the fish which swam across the screen. PF2 ported PF1 to a tablet and added some competitive elements by showing the most recent scores and allowing two subjects to play at once in a public space. PF3 added many more gamification features but the actual game play remained identical to that in PF1, except for the introduction of levels which changed the fish lifetime and inter-fish interval. In designing PF3 we considered the possibility of the changing avatars on different levels (e.g. using octopi or sea horses on different levels rather than the same fish avatar on every level). Another change we considered was moving to a 3D version with a realistic fish animation and underwater visual effects. We decided against these options out of concern that they might compromise the comparisons that were of primary interest to us.

Our exploration of gamification in the design of sensory neuroscience experiments suggests that gamification’s main advantages are

• if a tournament-style game is used, the experimenter can collect data more rapidly as dozens of subjects can participate simultaneously
• adding multiple levels to the game in the way we did, makes it possible to study the phenomenon of interest under different levels of pressure

The main disadvantages of gamification seem to be

• differences among subjects’ experiences introduce uncontrolled, confounding variables
• including multiple levels reduces the amount of data collected at any one level, which can diminish the statistical power of comparisons, within any level as well as between levels. - by discarding many of the constraints that would be imposed in a traditional experimental design, gamification expands the space of subjects’ possible behaviors. When exploratory data analysis is used to track subjects’ individual trajectories within this expanded space, the results, as we have seen, can reveal surprising strategies as well as potentially important differences among individual subjects (Turner et al 2017).

After considering the pro’s and con’s of gamification, we believe that for work in many areas of sensory and cognitive neuroscience research, the benefits of gamification make it a valuable tool to add to the experimental design process provided an appropriate balance is struck between the level of gamification and the experimental control over the subject’s experience within the game.

Acknowledgements

We thank Barbara Shinn-Cunningham, Boston University, for important contributions to early stages of this project.

Declaration of Conflicting Interest

The Authors declare that there is no conflict of interest.

Funding

Supported by CELEST, an NSF Science of Learning Center (SBE-0354378), supplemented by a grant from Brandeis University’s Office of the Provost.

References


