A Portal for the Online Evaluation of Serious Games

by

Velian T. Pandeliev

A thesis submitted in conformity with the requirements

for the degree of Master of Science

Graduate Department of Computer Science

University of Toronto

A Portal for the Online Evaluation of Serious Games

Velian T. Pandeliev

Master of Science, 2011

Department of Computer Science

University of Toronto

Abstract

Aging brings about not only physical, but also mental and cognitive changes and challenges. As our minds begin to slow down in middle age, we become more acutely aware of their functioning, and more concerned with maintaining our mental capacities. Recent findings in neuropsychology indicate that a combination of lifelong factors such as a cognitive commitment to one's career and hobbies, social engagement, higher education and diet contribute to an increased resilience to the adverse cognitive effects of neurodegenerative diseases such as Alzheimer's disease that often occur later in life. This phenomenon, dubbed 'cognitive reserve' by Prof. Y. Stern, seems to suggest that increased mental engagement prepares or 'trains' our brains to resist the effects of aging or disease. As a result, the last decade has seen dozens of companies offering training regimens that claim to improve cognition, boost mental capacities and stave off the effects of mental aging.

Most people tend to begin such regimens in middle age, when the effects of cognitive decline are first felt. However, it is still unclear whether the effects of lifelong engagement can be approximated or replicated by dedicated regimens begun later in life. At the same time, scientific validation of these regimens is a costly and difficult process typically relying upon in-person experimental protocols.

This paper presents an alternative to in-person lab testing: Tangra, a portal for conducting randomized controlled trial experiments online. Built in the Django web application framework, Tangra supports investigators in designing studies, populating them with participants, and managing their progress from initial consent through assessment and intervention to debriefing. Tangra also provides the participants themselves with an integrated interface where they can follow their progress in the study and participate in experimental sessions as required. The purpose of the portal is not restricted to mental fitness. Rather, it is to enable online experimentation in a variety of fields, reducing the logistical and financial burden on labs and widening investigators' potential base of study participants.

To demonstrate its viability as an experimental platform, Tangra was used to conduct a between-subjects mental fitness study comparing two games from a Canadian mental fitness company to two well-established board games. Fifteen participants recruited online were asked to complete ten 30-45 minute sessions over the course of a month. Their cognitive abilities were assessed before and after the intervention. Data from the intervention, as well as participant retention and motivation rates, were collected and used not to assess the effectiveness of the compared interventions in improving cognitive abilities, but rather to provide sufficient evidence that we have identified and addressed the key issues of conducting such experiments online.

Tangra is a first step towards an integrated online experimental portal. In our design, we have identified and attempted to implement the crucial elements of in-person experimentation in an online protocol. These elements begin with recruitment, screening and identity validation, cover ethics, consent and privacy concerns, and extend beyond intervention, assessment and instrumentation to monitoring confounding activity, motivating retention and providing feedback, ending in evaluating the persistence of an intervention's effects. While Tangra successfully implements some of these elements, future work is needed to further its progress towards a reliable and trusted online validation protocol.

Acknowledgements

This project was made possible through the guidance and expertise of many people. The author would like to sincerely thank (in alphabetical order):

Dr. Ronald M. Baecker, who supervised this project and provided mentorship and inspiration every step of the way.

Alison Benjamin, who enabled recruitment by establishing relationships with seniors' organizations.

Michael Cole for lending us the use of two of FitBrains' mental fitness tasks.

Liam Kaufman, who helped with the design and delivery of the first Tangra prototype.

Garry Ing for lending his extensive interaction, typography and UI design expertise to the Tangra interface, and for the Tangra logo.

Karyn Moffatt for some much-needed research methods help.

Joshua Steinerman, M.D., who helped define the parameters of the experiment and the study design.

Dr. Yaakov Stern, whose work in cognitive reserve spawned an industry, for lending some of his time and experience, as well as for the Set Switching task.

Chieh Schen Teng for being the relief experimenter for one of the participants, for suggesting the Shipley vocabulary task and the Selective Reminding task, and for trying his hardest to break the portal in the first few days.

Kevin Tonon for the implementation of Boggle and for frequently diving in to help during the author's first weeks in the deep end of the Django pool.

Dr. Khai Truong for agreeing to take on the job of second reader, and for his invaluable comments in the final stages of this paper.

Dr. Daniel Vogel for agreeing to impart his experience in designing and delivering software from an HCI perspective.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

A widely used method for conducting research with human participants in psychology and clinical research today is in-person experimentation in a controlled laboratory setting. It has many advantages, chief among them the ability to isolate the factor being studied and to eliminate confounds by having ultimate control over the environment. Most psychology departments have well-established mechanisms for recruitment that usually target undergraduate students and offer small monetary compensation or extra credit.

The in-person research paradigm has a number of notable downsides. Longer study durations, which usually require higher participant effort and commitment, make it harder to attract participants to come to a lab without offering additional incentives. Additionally, studies requiring large numbers of participants are limited by the availability of members of the chosen population in geographic proximity to the experimental site, and by their labs' throughput capacities. Unfortunately, certain study designs exhibit several of these limiting characteristics, making them difficult to perform this way, for instance studies targeting long-term effects that require months of commitment over multiple sessions on the part of the participant.

A rapidly emerging example that could benefit from an alternative experimental paradigm is the field of serious games research. Over the last ten years, the gaming industry has evolved to appeal to a wider and wider demographic: 38% of Canadian gamers are female, while the percentage of gamers over the age of 55 has doubled between 2005 and 2009 to 34%[1]. Video games are becoming a common mainstream pastime, their popularity partly due to the proliferation of the 'serious games' genre. A serious game is one that is not primarily geared towards entertainment, but has other aims such as education, physical fitness or improving and maintaining mental capabilities. When serious games are accompanied by a claim of a positive outcome of some sort (e.g., improved memory, better social skills, or mastery of a language), scientific validation of such claims becomes necessary.

The genre of serious games that introduces the most specific validation challenges is the field of mental fitness. Mental fitness games and interventions are often marketed as having the potential to improve memory, mathematical reasoning, focused attention, planning or executive function. What makes them special in the field of serious games is that unlike persuasive or educational games, mental fitness games border on medical interventions, and users look to them to delay, prevent or counteract a perceived cognitive decline later in life. This makes them a priority for scientific validation, while simultaneously making them one of the hardest genres to validate.

The current protocols employed by mental fitness intervention research largely follow the same experimental design model (e.g., Ball et al. 2002, Smith et al. 2009, Owen et al. 2010). First, the target cognitive abilities are assessed before the intervention to establish a baseline. Then, the intervention occurs, usually in repeated sessions, two to three times a week for 1-6 months. At this stage participants may be assigned to one of several experimental conditions if different interventions are being compared,

1 Entertainment Software Association of Canada: Essential Facts (2005, 2009, 2010)

or they may be assigned to an active or passive control condition. Finally their cognitive abilities are assessed again. A significant improvement over the baseline indicates a positive intervention outcome.

Many of these studies (e.g. Smith 2009) require participants to come into a controlled lab setting for assessments before and after the intervention, as well as for each intervention session. This ensures that their identity can be verified and that they can perform tasks without distractions in the presence of an experimenter. Although scientifically sound, these methods incur a large logistical and financial cost, both to the lab conducting the study and to the participants themselves; experimenter hours, travel to and from the lab, and a time commitment over months on end make these studies difficult to conduct in person. Additionally, since mental fitness games are geared towards users in middle age or older, the logistical burden of testing them is exacerbated by the lower mobility and willingness to travel of this population.

In recent years, the idea of conducting experiments online has become a viable alternative to in-person testing. Online experimentation provides a possible solution to some of the challenges with mental fitness validation, since these interventions typically require no specialized hardware (unlike eye-tracking or touch screen studies) and many of them are already deployed as Web-based tools. Unlike PC-based experimentation, conducting experiments online is OS-agnostic in principle and should require few changes to participants' computers. Additionally, participant pools would no longer be restricted to geographically close participants, enabling broader and more representative recruitment. Online experimentation also has the potential to greatly reduce the amount of resources required to conduct such a study, enabling participants to complete sessions at their own pace, from the comfort of their own homes, cutting down on experimenter hours and lab space.

However, this notion brings with it an additional set of challenges which have already been solved in in-person experimentation: ensuring recruitment from the desired population, screening for exclusion criteria and validating participants' identities, maintaining their privacy and the security of their data, monitoring and accounting for confounding activities, providing appropriate adherence incentives and the necessary feedback to motivate participants, and evaluating the long-term persistence of effects.

This paper details a framework for conducting mental fitness validation studies online. It begins by discussing related work and the general challenges to such a framework. It then presents a functional Web-based prototype, and describes the feasibility study conducted with said prototype. Finally, it presents the qualitative and quantitative results of the study and outlines guidelines and directions for future work.

The remainder of this document is structured as follows:

Chapter 2 describes related research in online experimentation and cognitive fitness. **Chapter 3** identifies the challenges that must be tackled to enable trustworthy online experimentation and discusses ways to address them. **Chapter 4** describes the technical aspects of the Web-based prototype, including its architecture and interface design. **Chapter 5** details the design of the pilot study conducted, **Chapter 6** presents the results of the experiment, and **Chapter 7** discusses the implications of these results and guidelines for further exploration.

2. Related Work

2.1. Cognitive Decline, Training, and Mental Fitness

Evidence from Alzheimer's Research

Alzheimer's is a troubling disorder as its causes are still largely unknown, it affects a large and growing percentage (half a million Canadians are afflicted, expected to grow to 1.1 million within 25 years[2]) of the population and its symptoms are severe: loss of independence, trouble recognizing loved ones and eventually a complete deterioration of cognitive function. However, epidemiologic observation (Katzman et al. 1989) suggests that it is possible to suffer the neural degeneration associated with Alzheimer's without exhibiting detectable symptoms of cognitive decline. What Katzman identified as "brain reserve" or the ability of a larger brain to resist degeneration longer, has in fact been shown to be dependent on more than brain size or synapse count. Rather, there are lifestyle factors which affect the manifestation of Alzheimer's symptoms and render individuals with cognitively stimulating careers, greater social engagement and higher education more resistant to the detrimental cognitive effects of neurodegenerative disorders (Stern 2006).

Cognitive Reserve

Dr. Yaakov Stern uses the term "cognitive reserve" to describe the mind's resilience and adaptability to brain degeneration. Several studies provide evidence for some of the factors associated with higher cognitive reserve. To control for lifestyle differences, many of these studies are conducted on monastic populations which have similar daily routines, vocational duties and diets. A study on a convent of elderly Catholic nuns (Snowdon 1989) determined that nuns with a Bachelor's degree were 2.5 times more likely to live to an advanced age and to retain the ability to care for themselves at that age than their counterparts who only completed grade school, indicating education as a factor in independent living and cognitive reserve. In another, longitudinal study of 801 older Catholic nuns, brothers and priests, Wilson et al. (2002) tracked participants' cognitive decline over a mean period of 4.5 years while collecting data about the frequency with which they engaged in cognitively stimulating activities such as reading magazines and playing games. Participants were assessed before and after the study on a range of cognitive functions. The study found evidence that more frequent participation in cognitive activities in daily life was associated with a reduction in the decline of cognitive functions by 30-60% . Additionally, fewer cognitively active participants were diagnosed with Alzheimer's disease during the study period.

Cognitive Training and Mental Fitness

Wilson's (2002) conclusions citing lifelong cognitive activity as a factor in the resilience to decline led to the idea that it may be possible to elicit comparable benefits through active cognitive training regimens. However, it is still unclear whether starting such a regimen at age 45 or 55 could match the positive effects of lifelong factors such as education, diet and cognitive activity. Furthermore, to have a chance of observing a measurable improvement as a result of training, mental fitness interventions should be fairly

[2] Alzheimer Society (http://www.alzheimer.ca/english/disease/stats-intro.htm)

intensive, involving daily or almost daily activity over a period of months if not years. This requirement determines the logistical parameters of validation studies of mental fitness regimens.

2.2. Serious Games

In a related development, over the last ten years the gaming industry has evolved to appeal to a wider and wider demographic. Video games are becoming a common mainstream pastime. The broader spectrum of gamers today has decoupled video games from the sole aim of entertainment. This is evidenced by the proliferation of ‘serious games’ with primary aims such as education, physical fitness or improving and retaining mental capabilities.

When such games are accompanied by a claim of a positive outcome (e.g., Russoniello 2009) of some sort (e.g., improved memory, combating depression, or mastery of a language), scientific validation of such claims becomes imperative. This is especially important in the mental fitness field, but it applies to any game, 'serious' or not, which aims for anything beyond entertainment.

Serious Games as Cognitive Trainers

With the advent of age certain cognitive functions experience a natural decline. As people begin to notice this decline in middle age, they begin to look for interventions to help them combat its effects. As a result, companies such as Nintendo, FitBrains[3] and Lumosity[4] are increasingly producing games that claim to provide the means for cognitive exercise or training. According to a 2010 market report by mental fitness market researcher SharpBrains[5], this sector of the serious gaming industry alone was valued at 265 million dollars in 2008, and that number is expected to grow to 2-8 billion by 2015. Given the intensity and length requirements for cognitive interventions, motivation and adherence emerge as important factors to consider when designing trainers. Games are therefore a suitable medium for cognitive training as they include an entertainment component, making it more likely for users to play them longer and remain engaged. From a marketing perspective, a game with benefits (like the Wii Fit or Nintendo’s BrainAge[6]) is an appealing notion, combating both the idea that playing video games is somehow detrimental as a pastime and the idea that learning or improving oneself is necessarily boring and tedious.

Expectations of Cognitive Trainers

Lifelong factors aside, the main question before cognitive trainers is: “Starting at 45, or at 55, or at 65, will any amount of regimented cognitive exercise have a positive effect on one’s cognitive reserve?” Many researchers believe that to be the case, and so end users are left with the impression that mental fitness trainers will help delay cognitive decline if used correctly. However, in this area, industry is pushing forward without appropriate scientific backing - dozens of trainers are already on the market

[3] www.fitbrains.com

[4] www.lumosity.com

[5] www.sharpbrains.com/executive-summary

[6] www.brainage.com

while the extent of the benefits they offer is largely unknown. This is due in part to the fact that gaming companies are not required to validate their products for cognitive benefits to be successful, and in part to the logistical challenges of traditional validation studies.

2.3. Validation of Beneficial Outcomes

The concept of scientific validation is fairly new to the gaming industry. While games are rigorously tested for stability, usability and engagement, they have traditionally not had to substantiate scientific claims to sell well. Serious games, on the other hand, often offer explicit beneficial outcomes that require validation.

For example, in the mental fitness field alone, a simple Internet search yields dozens of companies that offer mental fitness games. These games would greatly benefit from being able to make claims to objectively improve mental faculties, but only if such claims were supported using research methods known from experimental psychology. However, with a few exceptions such as appear in Ball et al. (2002) and Smith et al. (2009), most such claims have not been scientifically validated.

Dangers of Poor Validation

Poor, insufficient or deceptive validation is a nuisance and a false advertising risk in most fields. However, when it comes to serious games, especially mental fitness games, poor validation can cause actual damage to users. Cognitive training borders on medical intervention, to the extent that the study detailed in this paper was initially required to register as a clinical trial. Claiming to improve mental health with a product and failing to do so is as grave as claiming to improve physical health and failing to do so. When users purchase a cognitive intervention, they are trusting the chosen product with their mental abilities, relying on it to help improve or maintain them.

The danger of poor validation concerns not only ineffective interventions, but also ones that may actually cause cognitive damage. For example, Zimmerman et al. (2007) found that the viewing of video content by infants aged 8 to 16 months had a detrimental effect on their language development, prompting the Disney company to recall a line of Baby Einstein DVDs that had initially been marketed to improve language development (Lewin 2009).

The importance of validation is not lost to the creators of cognitive interventions, but their ability to validate is hampered by the reality of what such validation entails in terms of financial and logistical costs.

Current Validation Methods

While well-established and valid, most cognitive experimental protocols require a controlled setting and the presence of an experimenter to guide the participant through the experiment. The main reason for this is that cognitive assessments like the Wechsler Adult Intelligence Scale (Wechsler 1939), which have become the standard for reporting cognitive ability, are specifically designed to be administered in person. They are paper-and-pencil tests, standardized as such on large samples, and they are only available to qualified neuropsychologists since they require a great deal of background knowledge to administer and score. The companies behind them are selling trusted, well-validated products, and are understandably reluctant to pursue any transitions to the online medium, as those would require a complete redesign and additional investments to validate and normalize the test using these new procedures. This has left the mental fitness industry without access to trusted online assessment tools, which has in turn confined validation studies to the in-person model.

These face-to-face studies are costly and difficult to arrange and conduct. For instance, a recent ACTIVE study (Wolinsky et al. 2009) which used cognitive training to treat depression involved ten in-person one-hour sessions with over 2000 participants aged 65 and over. Participants completed the interventions in groups of 3-4, but even so the study required over 5000 hours with a highly trained and certified experimenter present. This incurred a huge cost in rented space, staff training, working hours and participant travel. Similarly, in an as-yet unpublished study involving an attention and goal-switching game called Space Fortress, 60 elderly participants commuted to Columbia University, sometimes for up to two hours a day each way, to participate in a one-hour lab session[7]. They did this three times a week for 12 weeks, spending up to 144 hours in transit for a 36-hour intervention.

Mental fitness training is likely to require much longer trial periods if its results are to match those obtained through lifelong activity, which means that validation studies for mental fitness regimens will require more sessions over a longer time period, exacerbating the logistical difficulties in scheduling and conducting them.

2.4. Online Experimentation

Online experimentation is decidedly not a novel concept. It is the paradigm of choice for surveys, questionnaires and many studies that do not require supervision or the aid of an experimenter. The chief benefit of online experimentation is a geographically unrestricted participant pool, which is especially important when evaluating small, sparsely distributed populations. Other benefits cited by participants include less pressure and scrutiny during the study, better control over scheduling and the elimination of the need to commute to a lab. For experimenters, it enables them to test larger samples by reducing the amount of hours spent supervising participants, eliminates scheduling difficulties and allows them to focus on troubleshooting.

Online Serious Game Validation

The ability to conduct a study entirely online, taking participants from recruitment through intervention to debriefing in the comfort of their own homes, would be decidedly helpful in alleviating some of the issues with serious game validation outlined above. Several companies have made strides towards online experimentation and assessment. Cambridge Brain Sciences, a spin-off of the UK's Medical Research Council, has developed a Flash-based battery of tasks in four cognitive domains (memory, reasoning, concentration and planning) that casual users can complete to obtain a general idea of their cognitive abilities relative to other users. San Francisco-based mental fitness company Lumos Labs has created a set of online brain training exercises and has collaborated with researchers at Stanford, UC Berkeley and University of Michigan to validate or develop new tasks. Unrelated to cognitive health, Bolt|Peters[8] has

[7] Yaakov Stern, personal communication

[8] boltpeters.com

conducted innovative research emphasizing the importance of ecological validity in game testing (Bolt 2008), conducting user experience testing on EA's Spore in simulated home environments.

While isolated strides are being made, few general guidelines have been gleaned and no general framework currently exists for conducting valid research online.

Large-Scale Online Experimentation: A Case Study

A recent online study by the UK's Medical Research Council Cognitive and Brain Sciences Unit (Owen et al. 2010) is one of the most ambitious large-scale online studies designed to validate mental fitness claims in the last few years, with a neuroscientist as a principal investigator and the resources of a government institute behind it. It was conducted in partnership with the British Broadcasting Corporation, which used a popular science program to broadcast a call for participation. The study followed a clinical trial model, testing participants' cognitive abilities before and after a 6-week intervention and involved no live contact with an experimenter. Participants were randomly assigned to one of three conditions. In the first experimental condition, participants were trained on reasoning, planning and problem-solving exercises designed to improve cognitive ability. The second experimental condition involved training on tests (rather than exercises) of memory, attention, visuospatial processing and arithmetic. Participants in the third (control) condition were asked obscure knowledge questions and had to find answers online. Cognitive assessment tasks were adapted from Cambridge Brain Sciences, a company that provides a small set of relatively well-validated tests of memory, planning, spatial reasoning and concentration. The study was designed to investigate whether benefits of training on a particular task would transfer to untrained tasks or improve general cognitive function.

Out of an initial sample of 52 617 participants aged 18-60 (Owen 2010), 11 430 completed the study, yielding a retention rate of approximately 22%. This rate was lower in the less engaging control condition (16%), where participants merely answered trivia questions. These rates provide a general idea of typical retention rates in large-scale online studies without any completion incentives. The procedure for the study cited a 10-minute requirement per intervention session, but did not report on the average training time over 6 weeks. Given that the average number of sessions was reported as around 25 sessions between the two experimental conditions, the total time spent on cognitive exercise could have been as little as 4 hours.

The MRC study's design also highlighted several unsolved problems with online experimentation in general. When the results of the study were published, seemingly finding no significant transfer of benefits gained by training on a set of mental fitness tasks to other, untrained tasks, they were immediately questioned (Fernandez 2010) on account of potential design issues such as the duration of the intervention and possible confounding factors due to the online nature of the study. Fernandez aggregated three principal objections to the study's conclusions.

A Time magazine article (Harrell 2010) outlined the first, quoting neuroscientist Torkel Klingberg, a proponent of brain training:

"The amount of training was low. Ours and others' research suggests that 8 to 12 hours of training on one specific test is needed to get a [general improvement in cognition]."

It is unclear whether Owen et al. were aware of the relatively low intervention time, but one must consider the potential detrimental effects that raising the minimum requirement to 20 or 30 minutes might have on retention in what are essentially casual, uncompensated volunteer participants.

The second objection comes from the same article, in which Klingberg argues that delivering the intervention over the Internet to participants' homes makes it impossible to control for distractions such as watching television or multitasking. This is a pitfall of online experimentation in general which currently has no easy solution. However, it is arguable that since most mental fitness interventions are often delivered in the form of online games and tasks, then validating them in such a way is ecologically valid as it closely approximates the conditions under which they would normally be used.

Finally, in an article in Nature (Katsnelson 2010) neurologist Peter Snyder, a critic of brain training, points out that the wide age range of the study's participants (18-60) is inconsistent with the population requirements of the research question. Older participants, who would typically be using mental fitness regimens, would have a higher variability and lower mean scores on pre-intervention assessments, Snyder argues, adding that the study would detect "a supernormal effect in a healthy person".

The older population was possibly underrepresented in the Owen study since older participants are harder to recruit for online interventions, as fewer of them are generally online and fewer of them are comfortable with the notion of online studies.

Overall, the criticism of the Owen et al. (2010) study illustrates that without an agreed-upon methodology for online experimentation, even the most scrupulously designed large-scale studies will stumble against unresolved issues in the field. This highlights the importance of a framework for online experimentation guidelines and best practices.

3. Elements of Online Experimentation

We are investigating the process of online experimentation from multiple angles, drawing on previous research outlined in Chapter 2 to generate an initial systematic exploration of the challenges involved in online participant studies. We have identified six general issues that must be addressed in any future guidelines for online experimentation. We discuss them below using examples and experience from our mental fitness research, but we believe them to be universal.

3.1. Recruitment, Screening and Authentication

Finding participants in a chosen demographic and ensuring they are suitable for the study is a challenge for any scientist. Online experimentation introduces a valuable tool in tackling this issue – social networking and electronic word of mouth could free recruitment from the burden of geographic proximity. Although challenges do exist when recruiting seniors online, they are expected to decrease in future years as Internet use by seniors in Canada is constantly growing (up 3.1 % since 2005 to 65.9% in 2009[9]).

Once a sample of participants has been gathered, they would need to undergo a screening process, where individuals with confounding medical conditions or those not in the chosen demographic are excluded. While some screening methods cannot easily be administered online, others, like the Telephone Interview for Cognitive Status cognitive screening questionnaire (de Jager et al. 2003) can be adapted for use over a desktop video conferencing platform.

An additional issue in online experimentation is the inherent anonymity and potential deceptiveness of an online presence. There is some danger of, and some motivation for, identity spoofing in online experiments: participants may ask a spouse or friend to complete a task they cannot complete or they may pretend to be eligible for a study they are not if its completion offers monetary incentives. Future work should focus on ways for experimenters to ensure that their participants are who they claim to be as they log on. This may include the use of webcams to take snapshots of participants as part of their authentication procedure or intermittently throughout the experimental session. Currently, live interaction (in-person or remote) appears to be the most reliable way to validate participants' identities, especially in studies with narrow inclusion criteria or those offering significant monetary incentives. These approaches all raise privacy concerns, which compound the issues outlined in 3.2. below.

3.2. Ethics, Consent and Privacy

With experimental mental fitness interventions, there can be a small measure of risk to participants' cognitive health. An example cited above is the recall of Baby Einstein DVDs by the Disney Corporation in response to new scientific evidence (Zimmerman 2007).

---

[9] Statistics Canada. Internet use by individuals, by selected frequency of use and age. http://www40.statcan.ca/l01/cst01/comm32a-eng.htm

Ensuring that a study is ethically sound comes with its own challenges. It is not clear whether university ethics review boards are equipped to correctly identify unethical procedures in proposed online studies, or that they are able to recommend revisions. Researchers, in turn, should be aware of the full range of ethics guidelines which they may have to abide by, such as requirements in Canadian universities that experimental data not be stored on non-Canadian servers.

Obtaining informed consent online is another challenge. Presenting a long, text-based informed consent form to a participant on a screen would be as ethically sound as doing the same in person, but online presentation is vulnerable to users skipping to the end and consenting without having fully read the form. This phenomenon is already prominent with software license agreements. Techniques such as a videoconference or phone briefing, or a wizard-like document that presents information in small chunks and tests knowledge at the end, could be used to ensure that users are well informed of the risks of each study. The advantage to the first approach is live communication, which allows participants to ask questions as they arise and gives the experimenter a chance to respond immediately.

Just as important as ethics is the issue of data security and privacy. As noted in 3.1., authenticating a participant, which is trivial in a lab setting, may involve storing images of the participant's face as part of the online procedure. This is not only a data management concern, but may also make participants reluctant to participate due to concern for their privacy. Irrespective of the format, extreme care must be taken with sensitive participant information and data – it should be safeguarded using a combination of methods, including server security, encryption and disassociating data from participant identities. Fortunately, many best practices from in-person experimentation are applicable here, like using key sheets as the only links between participant identities and data ad destroying them once a study is finished.

3.3. Intervention, Assessment and Instrumentation

For logistical and financial reasons, online experimentation is a great choice for longer interventions, especially those which would normally be deployed at participants' homes and completed on their timeline. Additionally Web technology allows for interventions to be thoroughly and unobtrusively instrumented to record participants' performance and interactions, and also can include an easy way for them to provide feedback as they encounter issues. The possibility for gathering rich, multifaceted data is already being used, for instance by Jimison et al. (2004) to detect cognitive decline by instrumenting interactions with input devices on non-related tasks.

Validation of mental fitness interventions is typically commissioned by a product vendor and occurs with their knowledge and support. However, that may not always be the case, and some vendors may view the uncertain outcome of a validation study as a potential risk. Therefore, a significant technical, as well as ethical challenge is the ability to monitor performance and assess interventions without having access to the source code of these interventions or consent from their vendor, especially in cases where unverified claims of improvement are being made.

Finally, to measure improvements after an intervention, some form of assessment is necessary. In some cases that may be relatively easy, for instance a Spanish vocabulary test at the end of a Spanish learning intervention. However, in the mental fitness field, assessments are typically long, comprehensive neurocognitive batteries, which measure a wide variety of cognitive skills (such as memory, planning, or visuo-spatial reasoning). These batteries have been validated over several decades with thousands of users, but they are largely pen-and-paper, and changing the administering procedure in as drastic a way as administering them online may mean restarting validation of the batteries themselves. An alternative is to build an online-specific set of assessment batteries, which Owen et al. (2010) have begun to do, but more work is needed to bring these new assessments to the same levels of adoption and validity as their pen-and-paper counterparts.

3.4. Confounding Activities and Distraction

To have confidence in one's findings, one should keep track of any other activities that participants may engage in that may influence the results of the study, both online and offline. This is especially true for cognitive interventions as participants may be engaging in a wide variety of other cognitively stimulating activities during the study period.

While monitoring confounding activities in participants' daily lives is a challenge for in-person and online experimentation alike, participants in a lab can at least be monitored for confounds or distraction during the intervention. In contrast, the danger of confounding activity during an experimental session is far more significant online. Participants completing studies in their homes without investigator supervision may be interrupted by family or visitors, snack, take unsolicited breaks, and even concurrently surf the Web.

Interestingly, encouraging work by Matthew Brehmer (2011) from the University of British Columbia found that demanding interruptions did not diminish performance on two online tasks, even in older adults. Brehmer had participants completing a self-administered cognitive assessment battery called C-TOC. At random intervals, demanding unrelated tasks would interrupt the primary task, requiring participants to complete them to return to the task at hand. Results from the study indicate that such interruptions did not significantly diminish performance on the cognitive tests for any age. This is initially promising, providing some evidence for the viability of allowing interruptions during cognitive tasks. However, monitoring confounding activities and distraction will still be a necessary aspect of online experimentation.

With respect to session-concurrent online activities, it is possible to use in-browser technology to log keystrokes and mouse movement for unobtrusive monitoring. A less detailed, but less invasive method of monitoring could involve logging the time intervals when a particular browser window is in focus. Finally, by sacrificing ease of installation and using existing desktop monitoring applications, detailed information can be logged about all computer-based activity, in-browser or not. In the extreme case, real life interruptions could possibly be monitored by taking pictures with the computer webcam at regular intervals.

All unobtrusive montiring raises further privacy concerns. Participants might be less willing to enroll in studies with higher levels of surveillance, and they should always be informed of the type of information being logged about their online or computer-based activity.

Confounding activities are not restricted to the intervention session itself. Rather, many cognitively stimulating hobbies and daily activities may affect the results of an intervention, whether it is delivered online or in person.

Monitoring such offline activities is much more challenging. It would have to rely on self-reports, day reconstructions or questionnaires – all methods that are participant-dependent and that have known drawbacks and inaccuracies. However, it may be possible to improve on the data collection process by making use of mobile devices and online reporting mechanisms.

3.5. Incentives and Feedback

Enjoyment is a significant factor of mental fitness games, as well as many other serious games, and it can have a huge impact on the success of a game. If an intervention is enjoyable, if it is played rather than worked at, there is a much higher chance that it will be used for a longer period of time, and there will be a higher chance of the user seeing its benefits. A crucial step, therefore, to generating and retaining a user base, is to make interventions fun and engaging.

However, enjoyment may prove to be insufficient in the verification stage. The casual user is under no obligation to play a game for any significant length of time, but a mental fitness validation study requires a significant amount of participation, which makes motivating users to commit to it a difficult task. While many studies involve only a few sessions and take a short amount of time to complete, the interventions we are exploring will require a significant commitment over several months. Convincing people to participate, and, more importantly, motivating them to continue participating until the study is done, are key challenges.

Online experimentation is very well suited to one-off experiments that involve an hour or two of activity and a simple monetary compensation. However, for longer studies, even with significant rewards, participant retention rates tend to be far lower than for in-person experiments. In a recent study, U.S.-based insurance company AllState offered a PC-based driving training program to a sample of its clients. Only half of the people who signed up and installed the software finished the study (Warden 2010).

Participant attrition could perhaps be minimized by utilizing one of the strengths of in-person research: scheduling regular (perhaps bi-weekly) live sessions with participants to ask them about their progress. We encountered some evidence in support of this technique, and discuss it in greater detail in section 7.1. The live interaction is less about data gathering and more about providing medium-term motivation for participants. It is likely to improve participation rates by making users accountable to a real person, and assuring them that their contribution is valuable and that the experimenter is interested in their progress. It is of value to the experimenter as well, as it allows her to gauge the level of participant motivation and respond to questions that may have arisen in the preceding two weeks.

Online studies are more poorly suited to traditional rewards like monetary compensation since there is some evidence (Mattila et al. 2003) that a significant percentage of mature users consider Internet financial transactions unsafe. This attitude is likely to have changed for the better since 2003, but providing personal information in order to receive online compensation may still be a barrier to adoption.

An additional challenge posed by offering monetary compensation is that it makes online studies a lucrative target for fraudulent enrollment by non-eligible participants and other scams.

On the bright side, the nature of online experimentation creates opportunities for new participant reward and appreciation mechanisms, such as donations in the participant's name, access to additional content and other virtual goods.

The Web also enables the use of novel motivation techniques. As an example, an economics study conducted in 2006 (Ashraf et al. 2006) reaffirmed the motivating potential of commitment contracts and led to the creation of stickK.com, a website where individuals can set goals, stake money or reputation on meeting those goals, and nominate a referee to track their progress. If they fail to meet their goals, their stake goes to a predetermined recipient, often an individual or group they do not wish to support. This approach has been shown to motivate participants to achieve their goals more than goal-setting alone, and we believe it can be modified to help provide the incentive necessary to complete interventions online.

3.6. Persistence of Effects

Although a particular mental fitness intervention may provide temporary improvements immediately after the study period, these improvements could be due to training effects and they could fade when the training stopped. For interventions claiming a lasting benefit, it is necessary to build follow-up assessments into the study to test whether these improvements are stable and transferrable. This is accomplished by conducting follow-up assessments some time after the end of the study, and measuring how well the improvements have been retained over the period of inactivity.

Another study from the ACTIVE clinical trial (Edwards et al. 2008) is a great example of long-term follow up – researchers conducted a follow-up assessment five years after the initial intervention. However, for most mental fitness studies the follow-up period is typically no more than three months, a timeline dictated by the academic deadlines of many social science researchers. These periods are insufficient to properly determine the long-term effects of interventions, and we believe that follow-up assessments should be conducted several years after the end of the study. This involves storing participant data longer, and introduces further motivation challenges. The benefits, however, are significant — interventions that can claim transferrable, lasting improvements would have a clear market advantage.

We believe that online experimentation can facilitate longer follow-up periods as participants can be reached anywhere in the world, digital data storage is inexpensive and both studies and follow-ups would be less logistically onerous on participants. On the other hand, follow-ups may be looking at granular, hard-to-detect changes which may be confounded by changes in the experimental environment since the initial assessment, including hardware changes.

4. Tangra: Architecture, Design and Technical Specifications

Named after an ancient Bulgarian deity, the Tangra web application is a prototype designed to represent a first pass at an integrated online experimentation portal. It supports investigators in designing studies, populating their groups with participants, and managing their progress from initial consent through assessment and intervention to debriefing. Tangra also provides the participants themselves with an integrated interface where they can follow their progress in the study and participate in experimental sessions as required.

The portal requirements were assessed over the course of several weeks in June 2010 under the guidance of Dr. Daniel Vogel with the help of Liam Kaufman. The design was refined over several iterations and essential features were implemented.

4.1. High-Level Organization

The Tangra portal was built to support a generic, field-agnostic experimental model. It recognizes two types of users: **Investigators** and **Participants**. Investigators create studies, which are hierarchical collections of certain high-level modules. Brief descriptions of these modules are detailed in Table 1.

At the core of the portal's organization is the notion of a **session**. A session is a singular activity or task that may take between 10 and 90 minutes to complete, such as signing an informed consent form, playing a mental fitness game, filling out a questionnaire or completing a debriefing interview. While only one informed consent session is expected for each study, other sessions may be repeated multiple times according to the requirements of the study. As such, we call a sequence of like sessions a **stage**. The term encapsulates a step in the completion of a study such as pre-assessment, intervention, or debriefing.

Participants see stages in sequential order. Depending on the study design, the order of stages may be different for different participants. Counterbalanced studies, for example, have a subset of participants completing various permutations of all intervention stages. Alternatively, not every participant may see every stage of the study: between-subjects studies routinely assign participants to one of several different conditions, which may consist of different stages. To enable investigators to control which stages participants see and in what order, stages are bundled into linear ordered sequences called **groups**. A group is roughly equivalent to a condition: each participant may only belong to one group and their membership determines what sequence of stages they will see. Finally, a **study** is a collection of participants split up into groups, either randomly or according to the study's requirements.

Term	*Definition*
Study	An experiment for participants to participate in and Investigators to manage that's broken down into sequential high-level steps (**Stages**)
Stage	A broad definition for one of the steps along the sequence to completing a **Study**. A **Stage** consists of one or more **Sessions**. A **Stage** can be a pre-study questionnaire, one of the interventions, or a bi-weekly interview.
Session	One of the sequential blocks of a **Stage**. A **Session** encapsulates a singular activity such as playing games for an hour, filling a questionnaire or conducting a phone interview. **Sessions** have deadlines associated with them. All **Sessions** in a **Stage** consist of the same activity.
Group	A defined sequence of **Stages** for a **Study** that a participant belongs to (follows). Examples of **Groups** are Treatment A/Treatment B or Experimental/Control. This is especially useful for counterbalanced designs: different participants who experience the same **Stages** in a different sequence would be assigned to different **Groups**.

Table 1: Glossary of terms used in Tangra high-level organization

Every participant starts the study in stage 1, session 1 of their respective group. Every time they complete a session, their session number advances. When they have completed the required number of sessions for each stage, they move on to the next one. This continues until they reach the last stage of their group, at which time they have completed the study.

This organizational model can support a wide range of study designs, including counterbalanced, A/B treatment and passive control study designs, in both between subjects and repeated measures studies.

4.2. Interface Design

Given the wide range of potential user populations, including children, the elderly and cognitively impaired users, the portal interface was designed first and foremost to be as simple and intuitive as possible for the participant.

From the architecture outlined in 4.1, it follows that every participant has a stage they are currently completing, as well as past and future stages, organized as a linear chain. The most important element among those, and the one featured most prominently in our design, was the current stage. The site was laid out so that the first thing every participant saw when he logged on was the current stage in the study they were involved in, the number of sessions already completed, the total number of sessions, and a large button which took them directly to the next session they were due to complete. Figure 1 shows an early low-fidelity mockup of this flow with the stages represented by squares in a row. Figure 2 is a screenshot from the final version of the portal, which has retained the left-to-right organization of stages, but includes better visual cues about progress.

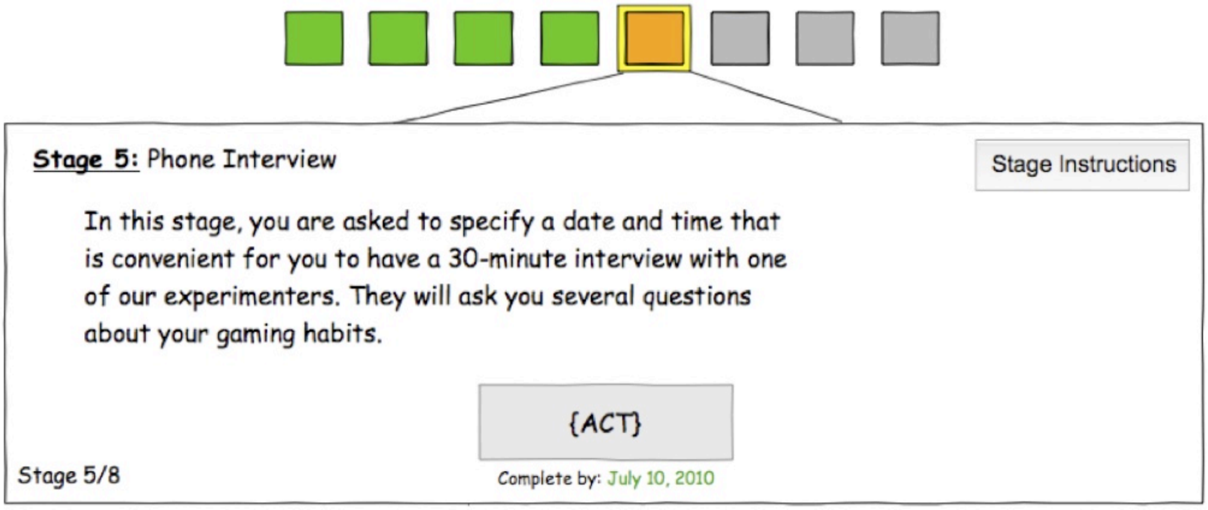

Figure 1: Early single study view mockup

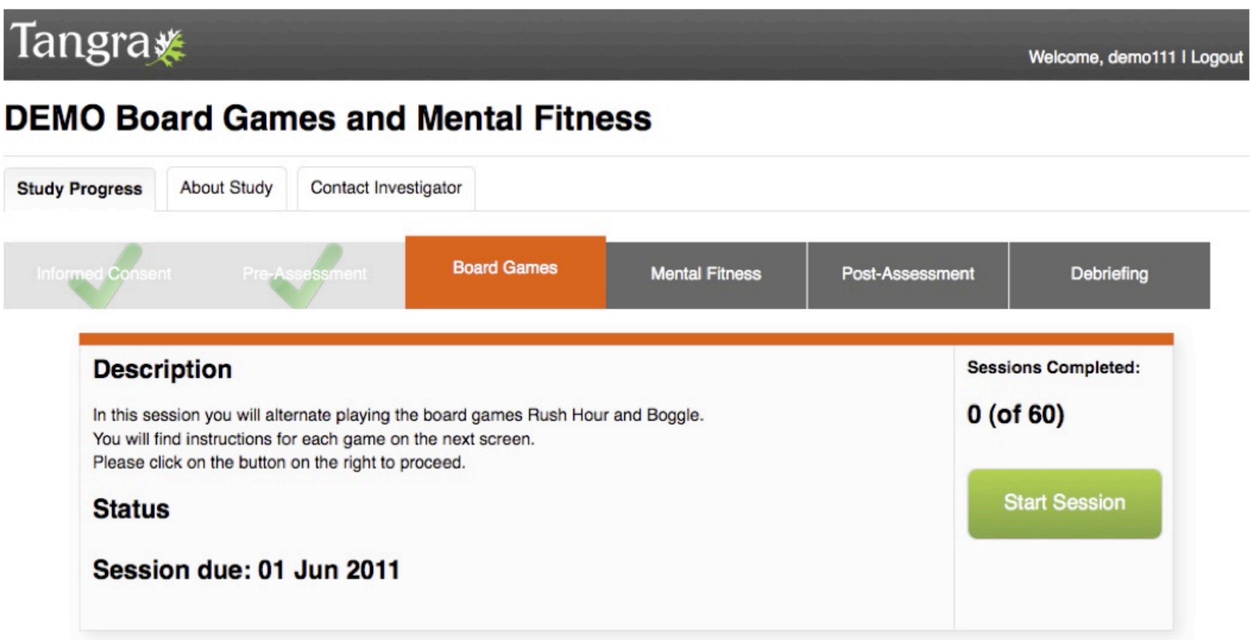

Figure 2: Final single study view

Participants could browse past or future stages by clicking on their titles in the horizontal bar, and there were additional, less visually prominent tabs for contacting the investigator or browsing study details. The most commonly required information and functionality was most prominently displayed. This enabled participants to begin a session of the study with a single click after logging on to the portal.

4.3. Implementation and Technical Requirements

The Tangra prototype was implemented using the Django[11] web application framework. Django is an open-source Python-based framework that comes with a web-based administrator interface and a secure user management module. The administrator interface allows authorized users to manipulate database contents from a graphical web interface.

High-level modules were implemented as database objects in Django's backend PostgreSQL database and were linked either through junction tables or through attributes. For instance, every stage object had a Study attribute denoting which study it was a part of. In contrast, to denote that a particular user was a member of a particular study, a Study-Participant object was created with two attributes: User and Study. This approach slightly increases database table complexity but provides great flexibility and the ability to denote many-to-many relationships between objects. The complete object relationship chart is shown in Figure 3.

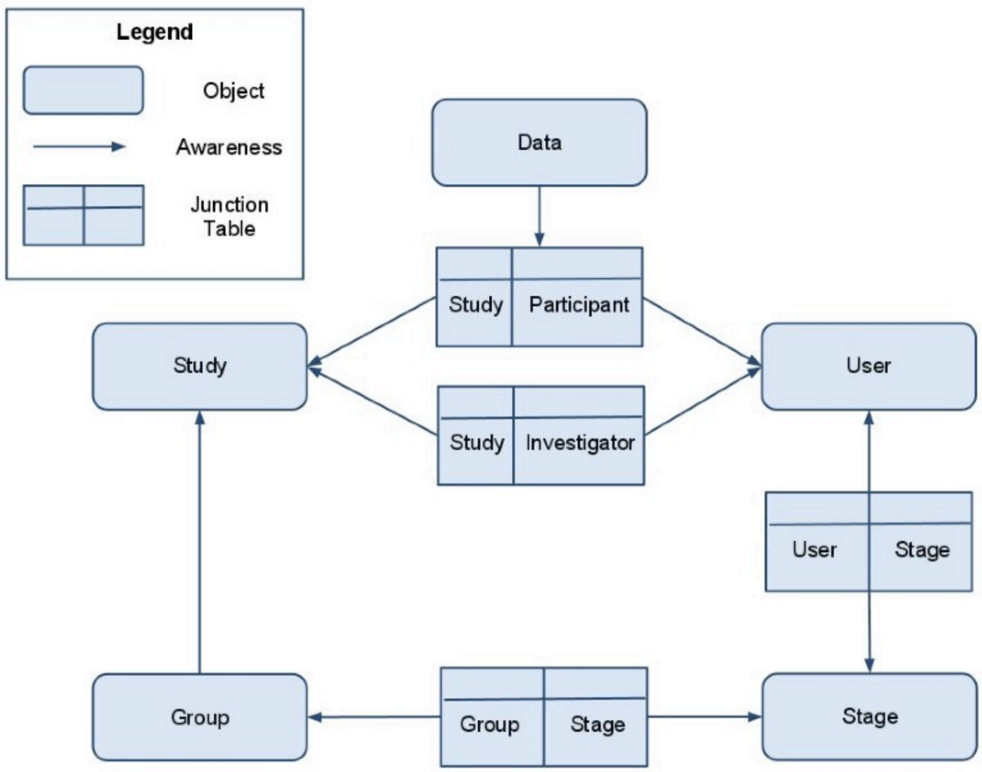

Figure 3: Database object relationships

[11] www.djangoproject.com

The front end of the portal was implemented in HTML and JavaScript (including extension libraries JQuery and Prototype). Due to Internet Explorer's poor compatibility with certain CSS attributes and JavaScript functionality, it was not supported for the prototype.

The Tangra portal was hosted on a dedicated server housed in the Department of Computer Science at the University of Toronto running Ubuntu 10.04.1 and served as an Apache 2 virtual host. The Django version was 1.2.5.

4.4. Tasks and Data Collection

Each datum the portal stores is a separate Data object in the database. It contains standard fields such as username, study and stage information and a timestamp, as well as an arbitrary datum field. Data objects are created by sending a POST request to a specific URL, which makes it possible to collect data from any page, both in-house or external, that can make such a request. In its simplest form, the POST request contains a single plain text string, referred to as the datum. On receiving such a request, Tangra automatically generates a received timestamp, user, study and stage codes, and creates a Data object to store the datum internally in an encrypted PostgreSQL database. The Data object contains all the relevant user and stage information, the Tangra-generated timestamp, and the datum as a plain text field.

The datum itself can contain a single measurement or piece of information, or it can consist of a sequence of comma-separated pieces. For example, a typical datum logged by an internal decision task may contain the response time, the response made, and whether the response was correct, all separated by commas. Data is exported in comma-separated text files, with the plain text field at the end. This makes it possible to store an arbitrary number of fields in each datum just by including commas in the datum sent. In fact, Tangra needs no special configuration to do this, and each POST request can store a different number of data pieces in a single datum.

Data objects may also be added manually to denote off-site activities or ones that the portal does not currently support.

While data collection is robust and easy to use, Tangra offers no data analysis tools. Fortunately, comma-separated data files can be analyzed by most statistical packages including SPSS.

4.5. Data Security

In compliance with Canadian data security guidelines, all data collected on the portal is stored on a secure in-house server, patched with the latest security fixes, in an encrypted server-side database. Additionally, the portal intentionally stores no personal information, identifying participants by unique participant numbers. To further protect participant identities, it is recommended that all correspondence between a participant number and identifiable information is confined to a single off-server electronic or paper key sheet, which can be permanently destroyed once the study is finished.

4.6. Implementing Elements of Online Experimentation

The Tangra experimental portal contains possible solutions to some of the six aspects of online experimentation outlined in Chapter 3. It takes all measures to secure participant data, collecting no identifiable information and conforming to social science best practices. It also uses an online-only informed consent process which would be a requirement for remote participants.

Most importantly, Tangra contains the core functionality for collecting data from specifically instrumented online tasks, both in the context of assessment of abilities and as part of an intervention with a conjectured beneficial outcome. By keeping consistent user profiles Tangra enables data collected at arbitrary intervals or by different studies to be associated with the same participant, facilitating the investigation of persistence of effects through follow-up assessments or studies.

Some aspects of online experimentation such as compensation mechanisms and recruitment are beyond the scope of an experimental portal and should be addressed in experimental design. Our pilot study (detailed in Chapter 5) provides examples of how these may be addressed orthogonal to Tangra's capabilities.

Other elements can be supported and enhanced by more advanced software solutions. For instance, Tangra implements a rudimentary authentication scheme based on a username and password, and performs no tracking of confounding activities during an intervention. More sophisticated authentication mechanisms like voice recognition or periodically taking photos of the participant could be implemented to increase confidence in participant identities. With respect to distraction and confounds, more elaborate activity monitoring would be required, both in-browser and on participants' machines, which, while unable to detect all confounds, would give a more detailed picture of computer-based distractions.

5. Study Design

5.1. Overview of Study

To demonstrate and assess the capabilities of the Tangra portal, a toy mental fitness study was designed to include as many elements of online experimentation as was feasible. The study compared two mental fitness tasks to clones of two standard board games that had a cognitive component, but did not explicitly claim to improve cognition.

Although the toy experiment is structured to approximate a clinical mental fitness trial, the goals of the overall validation study were not to make any claims about particular mental fitness games and their suitability to enhancing cognitive reserve, but rather to evaluate the portal's suitability to conducting such research.

5.2. Participants

For the first iteration of the portal, we set out to recruit 20 cognitively healthy adults aged 45 and over with English as their first language, computers at home and no sensory or cognitive impairments. We disseminated recruitment notices (see Appendix A) through several online publications, notably the Senior Alumni News of the University of Toronto and the online newsletter of the Canadian Association for Retired Persons. We reduced an initial pool of 34 potential participants who met the inclusion criteria to 15 on the basis of availability over the study period and technical requirements (specifically, use of or willingness to install the Mozilla Firefox Web browser). We issued participant numbers and portal credentials to those fifteen participants. Participants' ages ranged from 57 to 81 years old with a mean of 67.5 years.

5.3. Compensation

Drawing on studies of motivation and adherence in online contexts (Ashraf 2006, Gneezy 2010) and the perceived motivating factors of persons 45 and over, participants were offered no cash compensation. Instead, a donation would be made in their name to a charity of their choice, in the amount of 3$ for every session they completed (including cognitive assessments and interviews).

5.4. Experimental Design

The experiment employed a between-subjects 2x2 factor design: Intervention Type (Board Games vs Mental Fitness) x Contact (Skype vs No Skype). All participants signed an online informed consent form (see Appendix B), filled out a demographics and technology use questionnaire and completed two response-time-sensitive cognitive assessment tasks before and after the intervention. Additionally, participants in the Skype group completed two further assessments which required the virtual presence of an experimenter. Participants in the No Skype group had no contact with the experimenter before the study. All participants completed a debriefing interview, either over the phone (No-Skype group) or on Skype (Skype group).

The core of the intervention consisted of four intervention tasks: two mental fitness tasks which were closely matched to two board games on the basis of the cognitive domains they belonged to. A syllable matching task called Paradise Island was matched against the game of Boggle in the domains of verbal intelligence and spatial search while the traffic puzzle game Rush Hour corresponded to a delayed feedback spatial planning task called Wonder Juice Machine (see Table 2).

Every participant completed ten intervention sessions of approximately 45 minutes each over a period of 30 days. For participants in the Board Games condition, sessions alternated between Rush Hour and Boggle (even and odd sessions, respectively). Participants in the Mental Fitness condition alternated similarly between the two mental fitness tasks.

<table>
<tr><th></th><th colspan="4">Conditions</th></tr>
<tr><th rowspan="2">Session #</th><th colspan="2">Intervention Type</th><th colspan="2">Contact</th></tr>
<tr><th>Board Games</th><th>Mental Fitness</th><th>Skype</th><th>No Skype</th></tr>
<tr><td>1</td><td colspan="4">Informed Consent Stage</td></tr>
<tr><td>2</td><td colspan="4">Questionnaire Stage</td></tr>
<tr><td>3</td><td colspan="2" rowspan="2">Pre-Intervention Assessment Stage</td><td colspan="2">Online cognitive assessment tasks</td></tr>
<tr><td>4</td><td>(no task)</td><td>Live cognitive assessment tasks</td></tr>
<tr><td>5</td><td>Boggle</td><td>P.I.</td><td colspan="2" rowspan="10">Cognitive Intervention Stage</td></tr>
<tr><td>6</td><td>Rush Hour</td><td>W.J.M.</td></tr>
<tr><td>7</td><td>Boggle</td><td>P.I.</td></tr>
<tr><td>8</td><td>Rush Hour</td><td>W.J.M.</td></tr>
<tr><td>9</td><td>Boggle</td><td>P.I.</td></tr>
<tr><td>10</td><td>Rush Hour</td><td>W.J.M.</td></tr>
<tr><td>11</td><td>Boggle</td><td>P.I.</td></tr>
<tr><td>12</td><td>Rush Hour</td><td>W.J.M.</td></tr>
<tr><td>13</td><td>Boggle</td><td>P.I.</td></tr>
<tr><td>14</td><td>Rush Hour</td><td>W.J.M.</td></tr>
<tr><td>15</td><td colspan="2" rowspan="2">Post-Intervention Assessment Stage</td><td colspan="2">Online Cognitive Assessment Tasks</td></tr>
<tr><td>16</td><td>(no task)</td><td>Live cognitive assessment tasks</td></tr>
<tr><td>17</td><td colspan="2">Debriefing Stage</td><td>Semi-structured Skype Interview</td><td>Semi-structured phone interview</td></tr>
</table>

Table 2: Sessions and stages of the experiment by condition

The study was designed to employ a wide range of elements of online experimentation: response time tasks, interviews, live assessments, unsupervised online tasks and online questionnaires.

5.5. Tasks

This section contains brief descriptions of all tasks (assessment, intervention, questionnaire) used in the study. With the exception of the two mental fitness tasks, which were built in Flash and supplied through the cooperation of Canadian mental fitness vendor FitBrains, all online tasks were created in-house using JavaScript libraries. Questionnaires were administered through the LimeSurvey Web platform.

Demographics and Technology Questionnaire

The questionnaire consisted of three parts (see Appendix C). The first asked general information about the participant such as highest level of education attained, year of birth and marital status. The second dealt with hobbies and pastimes, gathering information on participants' physical, entertainment and creative activities. The third part dealt with computer usage history and habits.

Online Cognitive Assessments

Two tasks were chosen to test the portal's ability to record granular response time data (< 2000 ms). A digit/letter set switching task required participants to classify digits as even or odd and letters as vowels or consonants[12]. Participants also judged whether an arrow pointed left or right in the presence of confounding visual stimuli in a variant of the Flanker task (Eriksen & Schultz, 1974).

Live Cognitive Assessments

Two cognitive assessments which required the presence of an experimenter were administered remotely to participants in the Skype group (the group with access to video conferencing hardware). One was the Shipley Vocabulary Test (Shipley 1940): a verbal intelligence assessment which correlates highly with pre-morbid general intelligence in Alzheimer's patients. The other was the Selective Reminding Test (Buschke 1974), which provides a comprehensive score for working and long-term memory (see Appendix D).

Vocabulary Tasks

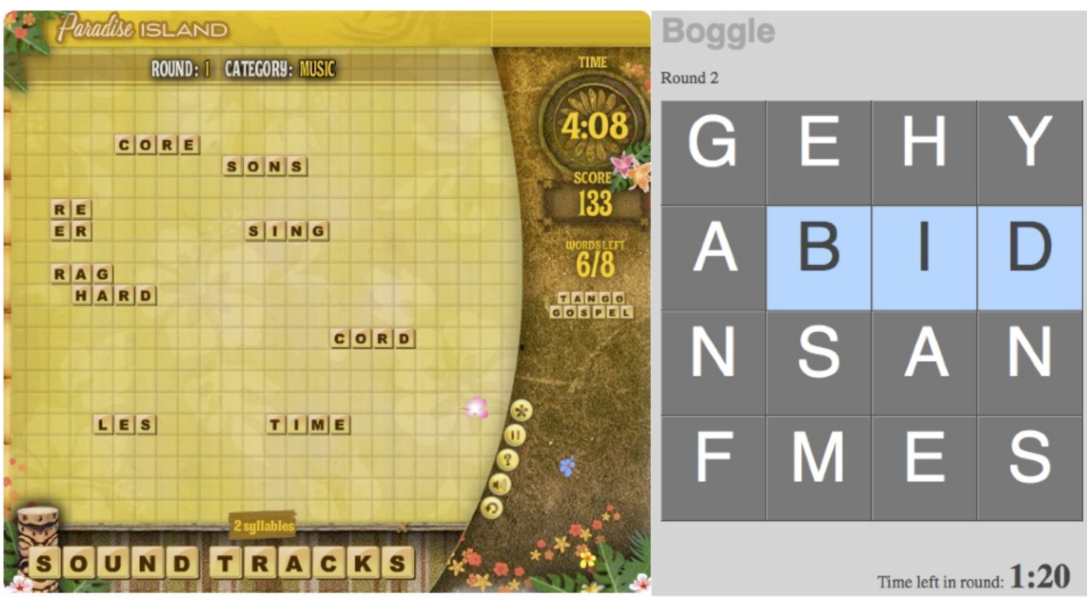

Figure 4: Paradise Island and Boggle

[12] This variant of the set switching task was designed at the Cognitive Neuroscience Division at Columbia University College of Physicians and Surgeons by Dr. Yaakov Stern and is as yet unpublished.

Two vocabulary and spatial search tasks were matched and assigned to the two intervention conditions (see Figure 4). The mental fitness game supplied by FitBrains was called Paradise Island and involved a combination of spatially matching syllables to form words in a particular category and rearranging Scrabble-like tiles into words. It was played by dragging items with the mouse. The classic board game that best matched it without incurring the wrath of Hasbro was Boggle: a 4x4 grid randomly populated by letters that participants combine horizontally, vertically or diagonally to create words, earning points based on their length. Participants could enter words using the keyboard, or click on consecutive squares with the mouse.

Spatial and Planning Tasks

Figure 5: Wonder Juice Machine and Rush Hour

The FitBrains spatial planning task, Wonder Juice Machine, involved dragging tiles from an inventory onto a 5x5 grid to direct differently coloured objects to different exit points. What made it interesting is that all the tiles had to be placed before the objects were set in motion, requiring a significant amount of strategy. Matched against it was the game Rush Hour which required dragging rectangular blocks with the mouse to clear a path out of the grid for a target red block, involving a similar amount of planning and mental simulation. When the path was cleared, the red block would slide out of the grid using JavaScript animation and the participant could proceed to the next level. Figure 5 shows screenshots of both tasks.

Debriefing Interview

The debriefing interview (Appendix F) was conducted over Skype or telephone and included questions about the usability of the portal's interface, the overall experience with the study, comparisons to in-person protocols, attitudes on online gaming and general feedback. For participants in the Skype group, the interview coincided with the post-intervention set of live cognitive assessments.

5.6. Procedure

Participants were e-mailed their login information and a link to the demographics questionnaire. They agreed to the informed consent online, filled out the questionnaire and proceeded with the assessments.

Every participant on the Skype group was contacted by e-mail and a time to have a videoconference was arranged. The experimenter started the conversation by reading the instructions for the SRT, then he read out a list of twelve words. The participant repeated those they remembered, in any order, and the experimenter read the items that were not recalled. This continued until the participant recalled all twelve words correctly on two consecutive trials or until the sixth trial was reached. For the next fifteen minutes, the experimenter engaged in semi-structured casual conversation with the participant, distracting them from the words remembered previously. Then, at the end of the 15-minute distraction period, the participant was asked to recall as many words from the list as they could. Afterwards, for every word they did not recall, they were presented with four words and were asked to recognize which of the four was on the list. Finally, in the presence of the experimenter, the participant was directed to an online version of the Shipley vocabulary test. The No-Skype group did not participate in either assessment.

The next stage of the study involved two response time judgment tasks. Upon pressing the Start Session button, participants were shown the instructions for the Set Switching task. They put the index fingers of either hand on the Z and M key on the keyboard and completed four 30-trial blocks alternating between digits (even/odd) and letters (vowels/consonants). Then, they completed four blocks with the stimuli mixed. Participants were encouraged to take a break and look away from the screen between blocks. In the next session, they completed eight 30-trial blocks of the Flanker task with the same mechanics, making judgments about whether an arrow was pointing left or right.

Three of the intervention tasks were time-limited, with a 45-minute countdown displayed on the page. Participants played Wonder Juice Machine, Paradise Island and Rush Hour until the time ran out, then the window dimmed and a “Return to Study” button appeared. The session was logged as completed when that button was pressed.

In the fourth intervention, Boggle, participants were free to use the keyboard or the mouse to enter words they could see on a 4x4 grid for ten consecutive 3-minute rounds, while the board was randomized for each round. They returned to the study screen after the 10^{th} round was finished.

Every session in the study was paired with a suggested deadline for its completion, displayed below the “Start Session” button. The deadline was 10 days for the first session (the informed consent) and 3 days for all others, including the ten sessions in the intervention stage. Deadlines were presented as visually prominent but optional planning aids, intending to keep participants going through sessions at a regular, measured pace. Missing the deadline had no negative consequences apart from the text turning red and informing participants that they were overdue to complete the task (see Figure 6).

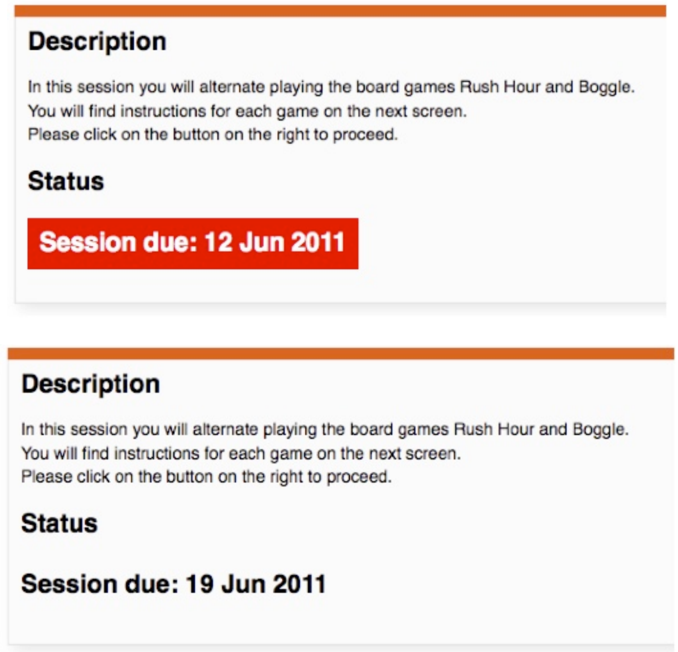

Figure 6: Overdue and non-overdue deadlines

5.7. Data Logging

The start and end time of every session was automatically logged server-side when the task finished. The online assessments logged a single datum for every trial which included the time taken to respond (as measured in the browser to eliminate the effects of network latency), the response, the expected response and whether the two matched. Rush Hour was instrumented to log the number of moves per level and the time taken to complete the level while Boggle kept track of every word entered, which round it was in, and how many points it earned.

To avoid redirecting to an external site and having a separate login to the FitBrains website, their games were embedded in an i-frame which did not interact with the system. Due to budgetary and time constraints, FitBrains server-side code was not altered to send data to the portal's server. Therefore, no data was logged for these games apart from the total session duration.

5.8. Research Hypotheses

Validity

The Tangra portal is reliable and precise enough to enable RCT studies to be conducted online and collects sufficiently granular data in an unobtrusive, ecologically valid way.

Viability

Using online methods to recruit participants, validate their identities and assess their cognitive performance will be as successful as employing the corresponding in-person methods.

Value

A study conducted online using the portal will have similar dropout and abandonment rates to in-person cognitive interventions of the same duration while being preferred by participants and incurring a far lower cost, both logistical and financial, to participants and researchers alike.

6. Results

This section presents the results obtained through the study. Quantitative results are presented first, followed by a qualitative analysis of the post-intervention interviews.

Out of 15 participants, 6 completed the study and one participant got as far as completing the post-intervention assessment but did not follow up with an interview. This additional participant is included in the analysis of quantitative session data, but not in the interview analysis or the retention rate calculations.

6.1. Quantitative Results

Participant Retention by Condition

The 15 participants were evenly and randomly distributed by Interaction Type (8 for Board Games, 7 for Mental Fitness), but only 4 participants met the hardware requirements to be added to the Skype group of the Contact condition. The other 11 participants were assigned to the No-Skype group.

Six participants (3 Skype and 3 Non-Skype) completed the study, yielding an overall retention rate of 40%. The Skype group had a higher retention rate (75%) than the No-Skype group (27%). The retention rate of the Board Games group was 5 participants or 62.5%, while only 1 participant (14%) finished in the Mental Fitness group.

Out of the nine participants who did not complete the study, three received their credentials and never logged on to the portal, one stopped after filling out the informed consent form, and two did not progress past completing the pre-intervention assessment. One participant, as mentioned, got to the end of the intervention, completed the post-assessment and failed to arrange a final interview. Finally, two participants withdrew from the study: one due to unrelated computer issues and one due to a spouse's health problem.

Cognitive Assessments

All participants completed a Flanker task and a Set Switching task before and after the intervention. We did not expect significant differences as the purpose of the intervention was not cognitive training in either task's domain. Rather, results here are reported as a testament to the portal's ability to obtain them.

We measured two variables for the Flanker task. The first was Response Time (ms), indicating how long it took participants to press a key after the stimulus was presented on the screen. The second was Accuracy (%) indicating the percentage of arrow direction judgments which were correct.

The seven participants who completed both the pre- and post-intervention assessments had an average RT of 612 ms before and 506 ms after the intervention. The difference was not significant ($t(6) = -1.64$, $p > 0.10$, $r^2 = 0.14$). Accuracy was 99.6% before and 97.4% after, also not significantly different ($t(6) = -1.16$, $p > 0.20$, $r^2 = 0.06$).

For the Set Switching task, we report only data from the last four trials for each test, the ones with mixed-format stimuli. We report the same measures as the Flanker task. Average pre-intervention RT was 837.1

ms, and post-intervention RT was 733.7 ms (difference was non-significant, $t(6) = -1.77$, $p > 0.10$, $r^2 = 0.18$). Pre-intervention accuracy was 98.1% while post-intervention accuracy was 98.7% (non-significant, $t(6) = 0.21$, $p > 0.50$, $r^2 = 0.001$).

The three participants in the Skype group completed two additional cognitive tests: the Selective Reminding Test (Buschke 1974) and the Shipley vocabulary test (Shipley 1940). These participants obtained 95% accuracy on the Shipley vocabulary test and scored 64/72 on average on the Immediate Recall section of the Selective Reminding Test. Delayed Recall (15 minutes after immediate recall) was on average 10.33/12 and Delayed Recognition was perfect (12/12) for all three participants.

Interventions

Similarly to the cognitive assessment data, results here are reported to demonstrate the granularity of the portal's data collection abilities. Additionally, since the two mental fitness tasks were accessed on an external website with no instrumentation, no data on the completion of these tasks is available beyond session duration.

Five participants played five sessions of Boggle each, consisting of ten 3-minute rounds. Due to an early deployment bug in the system, data for one participant was logged incorrectly and was subsequently excluded from the analysis. When analyzing Boggle data, it was determined that the open-source dictionary used was not rich enough and yielded many false negatives. Words that should have been accepted such as NICE, HAS, LENT, TAR and HARE were not recognized. This was corroborated by participant feedback (reported in detail in 6.2.) and contributed to participant frustration. With this limitation, we observed an average of 14.5 words per participant per round. The most complex words found by participants included BUSTLE, FROLIC, SHIVER, SIGNET and FOREVER. We encountered a small difference ($M = 38$, $SD = 62$) in word scores (see APPENDIX E for full scoring rules) between the first and the last Boggle session, indicating a possible training effect. The difference was determined to be non-significant ($t(3) = 1.55$, $p > 0.20$, $r^2 = 0.44$).

Rush Hour is a task consisting of increasingly difficult levels. Participants completed between 12 and 38 levels each (26.6 on average) over five 45-minute sessions. As expected, beginner levels took less time (134 seconds on average) than intermediate (250.5 s), advanced (287.3 s) or expert (335.6 s) levels.

Session Completion Patterns

Since no rigid scheduling instructions were given, participants were free to decide when to complete sessions. Through the use of session data, different pause intervals between sessions were plotted by frequency.

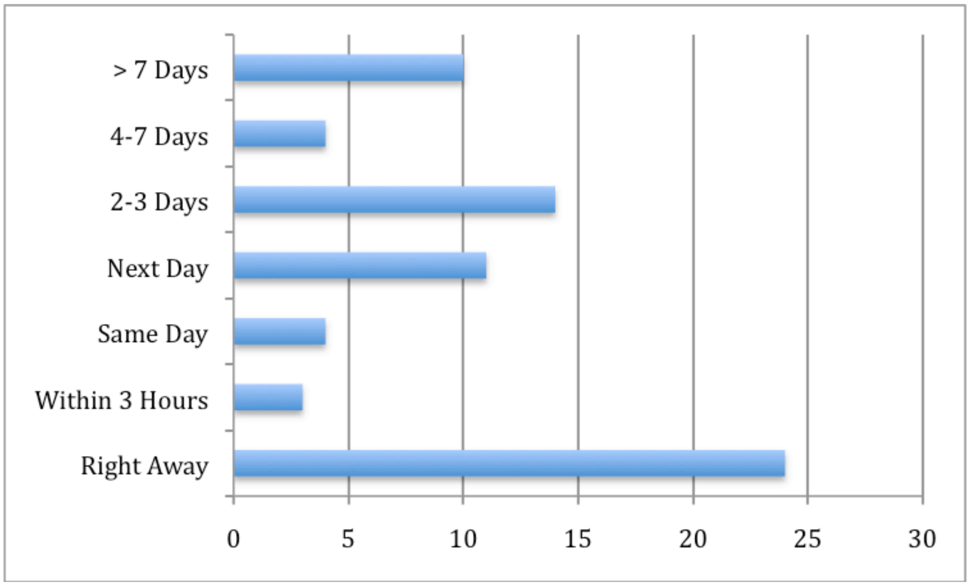

Figure 7: Session Completion Intervals

6.2. Qualitative Results

All the participants who finished the study underwent a semi-structured live interview with an experimenter. Due to our initial underestimation of the dropout rates for the study, the provision to interview participants who had intentionally withdrawn or abandoned the study was unfortunately not included in the study's ethics review. Three of the participants interviewed completed the interview over Skype, while the other three completed it over the telephone. The interviews were recorded with participants' verbal consent and analyzed. A brief summary of responses to the interview prompts and notable highlights is outlined below.

Adoption

All six participants affirmed that it was easy to learn how to use the experimental portal with no prior training and that they would use the portal again for a different study.

Additionally, four participants independently said that online experimentation is more convenient, and three mentioned that they would not participate in the same study if it was conducted in-person. In fact, only one participant indicated a willingness to consider participating in the in-person equivalent of the study.

Motivation

When asked what motivated them to persevere with the study, four participants cited the fact that they had made a commitment, four indicated entertainment, and two felt it was important to contribute to the scientific understanding of mental fitness. Additionally, four participants said that the study's reward scheme of contributing to charity was a welcome bonus, but did not serve to motivate participation by itself. As for what other schemes they would find motivating, one cited a cash reward and one indicated the desire to be entered into a draw for a prize. Overall, participants did not seem to be concerned with particular rewards, monetary or otherwise.

Attitudes Towards Online Gaming

Interestingly, only half of the participants asked indicated that they would like to continue playing the portal's games beyond the end of the study, and four indicated they were not generally interested in playing games online in the future. However, one participant said they had become addicted to Rush Hour and one said that Rush Hour was frustratingly challenging at times, which motivated them to try to beat it. Additionally, one mentioned that pen-and-paper activities such as crosswords are more convenient as they can be taken on a bus or to various areas of the house.

Scheduling

When it came to how participants scheduled their session completion, three said they casually logged on without following any rigid schedule, two followed the posted suggested guidelines, and one said they completed one session per day.

One said that the lack of consequences for missing a deadline was a good thing as it allowed for more flexibility, one said they could have used an e-mail reminder that they were due to complete a session, and one mentioned that they had completed four or five 30-45 minute sessions consecutively one day.

Issues and Improvement Suggestions

Among the largest issues participants encountered were the absence of feedback on the Set Switching task (3 participants), an incomplete Boggle dictionary (3 participants) and the lack of Internet Explorer support (2 participants). One mentioned that the portal was not entirely stable for the first few days after its release, one requested shorter sessions, and one found the ads that preceded the Mental Fitness group tasks confusing.

Experimenter Contact

All participants reported satisfaction with the level of experimenter contact during the study. Two praised the quick turnaround on e-mails with technical issues and one mentioned that the absence of an experimenter during the intervention reduced distraction and stress and made them feel less scrutinized. At the same time, one participant said she would prefer live communication when troubleshooting as it would be easier and faster.

7. Analysis, Discussion and Conclusion

This section discusses the results from the study, organized according to the main issues of importance identified by the study, followed by an overview of how the portal incorporated the six elements of online experimentation, and a response to the research hypotheses outlined in Section 5.8.

7.1. Issues of Importance

Recruitment Bias

Recruitment for the study was done through two seniors organizations: the Canadian Association for Retired Persons and the Senior Alumni Association at the University of Toronto. Calls for participation were placed with their respective newsletters, requiring participants to send an e-mail to an experimenter to participate. In fact, throughout the study they interacted with experimenters mostly by e-mail. This method of recruitment yielded a generally tech-savvy, cognitively high-functioning subset of seniors. Everyone who participated in this pilot study was either using Firefox already or had no trouble installing it. Additionally, participants in the Skype group scored very well on the Selective Reminding memory test and almost perfectly on the Shipley vocabulary test. This bias may result in smaller effect sizes for mental fitness interventions as training gains would likely be lower for high-functioning participants. Online recruitment is trivially impossible if the goal is to investigate seniors who do not own computers or do not surf the Web.

Third-Party Data Collection

Evaluating mental fitness and cognitive performance depends in part on ensuring that each task is being completed with due effort and attention, which typically requires that tasks be instrumented to monitor usage throughout. In this study, even though the overall time taken to complete third-party sessions was logged, no other data was gathered from third-party tasks because altering their code to send data to the Tangra server was deemed cost-prohibitive by our collaborating vendor. For larger studies, vendors may cooperate and invest the time and money to instrument their tasks to interface with Tangra, but they may also choose not to alter their code or not to cooperate at all.

The ability to gather data from third-party interventions without the cooperation or resource commitment of their vendors would make it possible to independently verify claims of improvement. This idea was briefly explored in the initial stages of the project and was deemed a useful feature for the next iteration of the portal. Many tasks have indicators of performance such as overall game score, number of items answered correctly, and overall time taken. Collecting this information would be essential to analyzing ongoing performance. Ideally, this would be accomplished by inserting POST requests to Tangra servers at appropriate places in third-party code to reliably provide researchers with the necessary information. We believe that this could also be accomplished without altering vendor code, albeit less reliably, through image processing. Task screens tend to have relevant information displayed in persistent visual areas. A screenshot could be taken of the task at regular intervals, and optical character recognition could be performed on areas that were identified as containing pertinent information. This technique does not require vendor cooperation and is much easier to apply to cooperating vendors as it does not require any changes to their product.

Quantitative Research

Valid, testable, time-sensitive data was obtained for two intervention tasks and two pre- and post-intervention assessment tasks. Session start and end times were recorded and it was shown that it is possible to instrument a Web portal to collect such data, including response time-sensitive data. This is not a new result, but it is a crucial first requirement to any online experimentation platform. Having accomplished this, Tangra can be used to help investigate the experimental process and to conduct further studies.

Online Experimentation and Gaming Attitudes

The attitude towards online experimentation from our participants was overwhelmingly positive, and many said they would not have been willing or able to participate had the study been conducted in-person. Deterrents such as the difficulty and cost of commuting, the weather and scheduling inflexibility were practically eliminated in the online study.

At the same time, participants really enjoyed live interaction when it was available. One mentioned specifically that "[i]t's nice to meet people face to face". When it came to troubleshooting, participants generally asked for help by e-mail and received timely responses. However, one participant noticed an experimenter's phone number on the informed consent form and called him at 10 pm for technical support. Given that this experimenter was a graduate student, 10 pm found him in the lab working on server stability issues, so the call was not an inconvenience. However, it does hint at a hypothesized tendency for people (and seniors in particular) to prefer live, synchronous and personal technical support if possible. Future iterations of the portal should include means of synchronous communication, either through a chat widget with an experimenter, or by using an embedded video conferencing platform.

Despite the positive response to online experimentation, something participants committed to in the interest of furthering science or the eventual cognitive benefits of such interventions, their attitudes towards online gaming were much more varied and not overwhelmingly positive. From our limited probing of this question, seniors are far from sold on online gaming. Some considered online gaming to be lacking the social interaction of in-person games, saying things like "staring at a computer screen is a lonely pastime". This is especially true for single-player online games. Additionally, many seniors do not regard online communication as particularly fulfilling and do not seek to play games online with friends and family.

The mechanics of online games are a further deterrent. One participant had actually suffered a sore bicep from repeatedly clicking on puzzle pieces in a previous online game. Another user said, "When playing Boggle, I preferred the mouse. Keyboards are harder for older generations to use quickly." This hints at the mouse being an easier input device to use by this population and stands in sharp contrast to power users, who prefer a heavier emphasis on keyboards and keyboard shortcuts.

Another important downside of online games is their lack of portability. One participant cited a love for crosswords, and spoke very highly of the ability to tuck a crossword and a pen into his shirt pocket and play it anywhere: on the subway, in a doctor's office, and in different rooms of the house. He saw online games as only playable at a desktop computer, and therefore being very restrictive. The idea of portable electronic gaming may not be appealing or viable for many seniors who either do not have the technology or do not see the need to adopt it. Online gaming portals in turn are not generally geared towards easy adoption by less tech-savvy populations, which compounds the problem and delays attitude changes.

Visual Elements and Design

The portal design contributed greatly to the positive attitude towards the platform, with all participants finding it very easy to learn, navigate and use, and perfectly acceptable for use in future studies. This is a significant success for the project, and underlines the importance of emphasizing clean, straightforward user interfaces. As Shim (2010) posited, a huge deterrent to online game adoption for seniors is a cluttered and intimidating interface.

The most important design principle followed in the portal design was minimizing the number of clicks required to bring participants to the next intervention session. The portal kept participants logged-on for up to two weeks, taking them to a page that showed their progress at a single glance while keeping the most important information (the current stage of the study and the big green button leading to the next session) in the most prominent place on the screen. Persistence of UI elements, lack of distracting animation effects and high-contrast, clear fonts contributed to the usability of the portal.

A significant challenge to the development of online content is browser compatibility. The biggest issue encountered when implementing Tangra is the lack of adoption of certain CSS elements by earlier versions of Internet Explorer. Future studies would be less hindered by this factor as Internet Explorer advances to match other browsers' capabilities.

Retention and Rewards

While the overall dropout rate (60%) was not unexpected, given the online nature of the study and the time commitment it required, it was slightly disappointing. However, the dropout rate was far higher (73%) in the No Skype group than in the Skype group (25%). Both groups were balanced across the same interventions and both groups were interviewed by an experimenter after the study. The crucial difference was that the Skype group had a live interaction with an experimenter before the intervention began. Despite the relatively low number of participants in this pilot study, this seems indicative of the importance of initial in-person contact with an experimenter for retention. Additionally, many of the participants asked what motivated them to continue with the study indicated a sense of commitment or duty, which can only be strengthened by a live interaction with an experimenter.

A study investigating the effect on retention of live interaction before a one-month intervention is planned for the near future to determine which element of the initial interaction was responsible for higher retention. Was it the live interaction itself, was it some content that could be delivered through other means (such as pre-recorded video), or were there other contributing factors? For the study, Tangra will be enhanced with an in-browser audio and video conferencing platform to eliminate the need for using third-party desktop applications. Participants without the necessary hardware for online audio communication will be contacted by phone. Overall, given the results of this study, it is likely that initial contact will increase retention, as will regular contact throughout longer intervention studies.

It was decided to compensate participants by donating to a charity of their choice, a decision driven by an urge to avoid participants outside the target population from signing for monetary gain. Additionally, it was believed that 45$ in cash would be less motivating to a retiree than it would be to a first-year undergraduate, and less motivating to a retiree than the prospect of contributing to science, the community, and making a charitable impact. This theory was not entirely disproven, with everyone enjoying the ability to make a contribution, but at the same time the compensation was not cited as the primary motivating factor by any of the participants we interviewed. One was even reticent to donate due to her previous experience with charity organizations hounding her for more donations once she had contributed once. Other participants mentioned they would have preferred to be entered into a draw for a larger prize.

Overall, online experimentation opens doors to novel reward mechanisms, but further work is needed to determine what combination of rewards would be the most motivating. Additionally, it became clear that the Department of Computer Science at the University of Toronto does not have the financial flexibility in place to support novel reward mechanisms. Currently, the only supported mechanism is in-person cash compensation, and the Department requires that participants provide their name and contact information, as well as sign a receipt in person stating that they have received the money in order to be compensated.

Engagement and Enjoyment

Enjoyment has enormous motivating potential. Games such as World of Warcraft have garnered enormous, compulsive followings without offering any benefits or rewards apart from in-game achievement, enjoyment, and a degree of social renown. Mental fitness products have very often taken the form of games in the hopes of making interventions fun and enjoyable alongside any cognitive benefits. In fact, some vendors frame their products as games for fun with a small cognitive component. Understanding what part enjoyment plays in motivation is a key factor in furthering online experimentation research.

Interviews with participants yielded a few interesting insights into how game engagement contributed to motivation while interacting with the portal. One participant said:

> “I was on the verge of being addicted to Rush Hour: on the one hand, it was extremely frustrating at times, but on the other, the animation of the car leaving the board at the end of every level was very satisfying.”

In fact, the Rush Hour task was the only one that adapted to participant progress. Third-party tasks had no way of knowing how far participants had previously gone, so they restarted from the very beginning for every session, and Boggle randomized the letter grid every time, with difficulty being largely a function of that randomization and not deliberately advanced by the game designer. Rush Hour, on the other hand, had a linear progression of levels which got more difficult as the game progressed. We found that when participants reached the first achievement plateau where a level was truly difficult for them, they reported a great sense of challenge, which motivated them to keep trying. Rush Hour also included a tiny reward: the animation of the red block reaching its target served to positively reinforce the sense of achievement at the end of every level.

At the same time, when participants truly reached a level of difficulty they could not surmount, motivation turned to frustration. One participant reported that when the Rush Hour level that he knew he could not solve came up, he would leave his browser running and step away, returning 45 minutes later to play the next session of Boggle. This is a serious issue with game design as it would reduce engagement and, by extension, the benefits of cognitive training. Therefore, a mechanism for adapting to participant skill levels should be a part of every long-term intervention.

Another interesting result came out of participant interviews. A participant reported that he did not rush through the intervention sessions because he enjoyed the anticipation of playing the games and rushing through them would have deprived him of that feeling. It is interesting to consider whether the same would be true if the participant could play these games an unrestricted number of times.

Finally, in contrast to the idea of enjoyment motivating benefits, one participant cited the fact that the games she was playing were considered good for her cognitive health as a justification or excuse for playing them zealously. This is an example of the beneficial nature of the activity having an impact on motivation, even if it was not a key factor in enjoyment. This is often observed with serious games, and may encourage people to stick with a game because of its benefits as well as its entertainment value.

Scheduling

Analysis of session scheduling showed a wide variety of time taken between intervention sessions. Some participants were diligent and completed sessions once every three days, as indicated by the due dates on the portal, while others fluctuated more. One reported having forgotten about the study and requested that an e-mail reminding system be implemented in the next iteration of the portal. It was also determined that, distressingly, a large number of sessions had been done one after the other on the same day. This is not advisable with mental fitness interventions as fatigue can diminish the effect of cognitive training and skew results, and hints at a serious issue with unrestricted scheduling.

A typical intervention has a recommended duration and frequency of activity, which is designed to include appropriate rest periods. An easy way to reduce consecutive session completion is to impose a minimum rest timeout after each session, making the next one unavailable in the meantime. On the other hand, one participant explained that she had rushed through a few sessions because she had been away for two weeks in the middle of the intervention, and another found the 45-minute interventions too long.

A scheduling scheme is easy to implement in in-person experimentation, but it leads to a methodological conundrum for online studies. Imposing more rigid scheduling and session duration requirements would ensure even delivery of the intervention, but it would be hard to enforce, and it would reduce the scheduling flexibility that participants greatly enjoy. Participants would be unable to complete a session if they were not certain at the start that they could finish it uninterrupted, and they would see a study as a greater imposition on their lives. Furthermore, it is not a realistic depiction of how mental fitness games are played in the real world. Proving that playing a game rigidly for 45 minutes every two days for six months has benefits is a step in the right direction, but it does not generalize to the fluctuating gaming habits of casual players. They may take vacations, have good and bad days, and play for different lengths of time, at different times, with different levels of concentration.

The main question to answer, a question beyond the scope of this project, is which scheduling method is preferable for which types of online experimentation.

Online Experimentation: Summary

Table 2 below summarizes the general advantages and disadvantages of online experimentation in addressing each of the six elements discussed in Chapter 3.

Element	*Advantages*	*Disadvantages*
Recruitment, Screening and Authentication	Recruitment is not geographically restricted and can be done on a large scale with minimal cost	Verifying that participants meet inclusion criteria is more difficult, as is validating their identities every time they sign in
Ethics, Consent and Privacy	Online experimentation is less invasive to participants and consent can easily be granted electronically	Potentially identifiable information may be sent along unsecure connections and may be susceptible to cyber attacks
Intervention, Assessment and Instrumentation	It is easier and more convenient to instrument and deliver tasks, with minimal experimenter involvement during intervention	Online assessment tools, especially for cognitive performance, are still in their infancy
Confounding Activities and Distraction	Online testing of interventions will more closely match the real-life conditions under which they are used	Accounting for offline confounding activities is just as difficult, and confounds may be introduced during the intervention itself
Incentives and Feedback	New mechanisms, online motivation strategies and non-monetary virtual rewards may be more motivating than traditional compensation	It is unclear what motivates participants in different demographics, and university administrations are rarely set up to allow for alternative reward mechanisms
Persistence of Effects	Follow-ups, even years later, are unaffected by participant relocation and participants are more likely to commit to them due to the lower logistical cost	Experimenters have less control over changes in the experimental environment, including participant-side hardware, which may influence follow-up results

Table 3: Advantages and disadvantages of online experimentation

7.2. Conclusions

Validity

The Tangra portal reliably collected detailed quantitative data, including precise response time measurements, without hindering user experience. This data was used to analyze login patterns and to demonstrate the portal's scientific potential.

Viability

Participants were successfully recruited online and several cognitive assessment tasks were administered, including well-established live assessment tests such as the SRT and the Shipley vocabulary test. Were it not for the vacuum created by the lack of well-established online assessments, Tangra could have been used to accurately and with confidence determined comprehensive levels of mental fitness. No issues of identity spoofing were encountered, yet future iterations of the portal will include means of reliably validating participant identities, possibly through facial logging.

Value

Participants completed a four-week intervention encountering no significant issues. Experimenter hour requirements were restricted to conducting several hours of live interviews, enabling more time to be spent on development and troubleshooting while maintaining high levels of participant satisfaction with experimenter contact. Participants were also overwhelmingly in favour of online experimentation as it reduced financial, logistical and scheduling burdens on them. The overall retention rate of 40% is relatively low compared to in-person interventions, but higher than other online studies have reported (Owen 2010) and valuable insights into increasing retention were gleaned during the pilot study.

7.3. Final Words and Suggested Next Steps

Although a proof-of-concept, the study conducted with Tangra is not categorically different from how a mental fitness vendor might evaluate one of its games. In fact, the number and arrangement of sessions could be very similar. To evaluate a game using Tangra, a researcher would have to find an appropriate assessment battery to implement before and after the intervention, and replace our intervention tasks with hers. Scaling the portal from twenty to two hundred participants would be a matter of using an optimized production-quality server. Although far from a complete solution, we believe this portal could be used to conduct a variety of real-world studies and could deliver trustworthy data.

Generally, there are several directions that future work on online experimentation could take. As an initial next step, in the near future Tangra will be used to investigate the effects of live communication on retention, after which it will be released as an open-source system. Additionally, an implementation of an in-browser video conferencing platform is already being tested to eliminate the need for participants to install third-party applications such as Skype.

A useful piece of this space would be a customizable image processing module which can take regular screenshots to determine scores and participant performance by OCR alone, without interfacing with a task's back end. This would reduce vendor-side costs for integrating with Tangra for testing, and would in fact enable third-party evaluation of interventions, for which vendor cooperation is optional.

Most importantly, to advance the field of online cognitive assessment and intervention, the vacuum of trusted online assessment tools should be filled with validated tasks designed specifically for Web-based use which can garner the same confidence and respect as in-person assessments such as the PPVT and the WAIS.

Tangra was created as a proof-of-concept online experimental portal. While time constraints made it prohibitive to conduct follow-up experiments, a cognitive intervention study was instrumented and delivered securely, logging detailed data and displaying a reasonable retention rate. Participants endorsed the portal through their willingness to use it for future studies and the ease with which they adapted to its interface. Useful guidelines for online experimentation were established through interviews and data analysis. It is an important initial step towards a robust general protocol for online experimentation.

8. References

Ashraf, N., Karlan, D., and Yin, W. 2006. Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines. *Quarterly Journal of Economics, 121* (2). 635-672. http://www.mitpressjournals.org/doi/pdfplus/10.1162/qjec.2006.121.2.635

Ball, K., Berch, D., Helmers, K. et al. 2002. Effect of cognitive training interventions with older adults. *JAMA 288(18)*. 2271-81. DOI=10.1001/jama.296.23.2805

Brehmer, M. 2011. On the Design of C-TOC: Self-Administered Cognitive Testing on a Computer. Master's thesis (in progress), University of British Columbia.

Bolt, N. 2008. Researching Spore: A New Approach to Player Testing. Presented at the Meaningful Play Conference, Oct. 10, 2008.

Buschke, H. & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology*, 24, 1019-1025.

de Jager, C.A., Budge, M.M. and Clarke, R. April 2003. Utility of TICS-M for the assessment of cognitive function in older adults. *Int J Geriatr Psychiatry, 18* (4). 318-24.

Delinski, B. 2009. Doctors link various injuries to Wii game system. Retrieved on March 11, 2010 from the Times Daily website. http://www.timesdaily.com/article/20090127/ARTICLES/901270328?Title=Doctors-link-various-injuries-to-Wii-game-system

Edwards, J. D., Ross, L. A., Ackerman, M. L., et al. 2008. Longitudinal predictors of driving cessation among older adults from the ACTIVE clinical trial. *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences 63(1).* 6-12

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics* 25 (4): 249–263.

Fernandez, A. 2010. BBC "Brain Training" Experiment: the Good, the Bad, the Ugly. SharpBrains website. Retrieved on Apr. 28, 2010 from http://www.sharpbrains.com/blog/2010/04/20/bbc-brain-training-experiment-the-good-the-bad-the-ugly/

Gneezy, A., Gneezy, U., Nelson, L. D. & Brown, A. 2010. Shared Social Responsibility: A Field Experiment in Pay-What-You-Want Pricing and Charitable Giving. *Science* 329, 325. DOI: 10.1126/science.1186744

Harrell, E. 2010. Study: Brain Exercises Don't Improve Cognition. *Time magazine*, April 20, 2010.

Jimison, H., Pavel, M., McKanna, J., and Pavel, J. 2004. Unobtrusive Monitoring of Computer Interactions to Detect Cognitive Status in Elders. *IEEE Transactions on Information Technology in Biomedicine*, 8(3).

Katsnelson, A. 2010. No gain from brain training. *Nature* 464, 1111 (2010). DOI: 10.1038/4641111a

Katzman R, Aronson M, Fuld P, et al. Development of dementing illnesses in an 80-year-old volunteer cohort. *Annals of Neurology*. 1989;25: 317–324.

Lewin, T. 2009. No Einstein in Your Crib? Get a Refund. The New York Times. October 24, 2009. Retrieved June 8, 2011 from http://www.nytimes.com/2009/10/24/education/24baby.html?_r=1&em

Mattila, M., Karjaluoto, H., & Pento, T. 2003. Internet banking adoption among mature customers: early majority or luggards? Journal of Services Marketing 17(5). 514-528.

Owen, A. et al. 2010. Putting Brain Training to the Test. *Nature* 2010. DOI=10.1038/nature09042

Russoniello, C., O'Brien, K. and Parks, J. M. 2009. The effectiveness of casual video games in improving mood and decreasing stress. *Journal of Cyber Therapy & Rehabilitation* 2(1). 53-66

Shim, N. 2010. A study of social interaction in online games for seniors. Master's thesis, University of Toronto.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, 9, 371-377.

Smith, G.E., Housen, P., Yaffe, K. et al. 2009. A cognitive training program based on principles of brain plasticity: results from the IMPACT study. *Journal of the American Geriatrics Society 57* (4). 594-603

Snowdon DA, Ostwald SK, Kane RL. Education, survival and independence in elderly Catholic sisters, 1936-1988. *American Journal of Epidemiology*. 1989;130:999–1012.

Stern, Y. 2006. Cognitive Reserve and Alzheimer Disease. *Alzheimer Disease and Associated Disorders*, 20(2).

Warden, T. 2010. Cognitive Training's Role in Driver Safety. SharpBrains Virtual Summit, January 2010.

Wechsler, David (1939). The Measurement of Adult Intelligence. Baltimore (MD): Williams & Witkins.

Wilson R.S., Mendes De Leon C.F., Barnes L.L., et al. Participation in cognitively stimulating activities and risk of incident Alzheimer disease. *Journal of the American Medical Association*. 2002;287:742–748.

Wolinsky, F.D. , Mahncke, H.W. , Weg, M.W.V., et al.(2009). The ACTIVE cognitive training interventions and the onset of and recovery from suspected clinical depression. *Journal of Gerontology: Psychological Sciences*, 64B(5), 577–585. DOI=10.1093/geronb/gbp061.

Zimmerman, F.J., Christakis, D. A., Meltzoff, A. N. 2007. Associations between media viewing and language development in children under age 2 years. *The Journal of Pediatrics 151* (4). 364-368. DOI=10.1016/j.jpeds.2007.04.

Appendix A: Recruitment Poster

Are you 45 or older?

Do you want to stay mentally fit?

The Technology for Aging Gracefully Lab at the University of Toronto is looking for participants aged 45+ who:

- Are healthy and have normal or corrected-to-normal vision
- Are familiar with computers
- Have access to the Internet
- Suffer from no cognitive impairments

The Study:

There are many products out there claiming that they can help delay the decline in memory and other mental functions that comes with age. But how useful are they really?

We are testing a new way to evaluate these products over the Internet.

Your role:

You will be asked to play some online games for ten 45-minute sessions in the comfort of your own home within a one-month period. We will then test to see if those games had a positive impact on your mental fitness.

Your reward:

In addition to contributing to our understanding of mental fitness and playing some fun online games, you have a chance to help others in a more direct way: we will donate 3$ to the charity of your choice for every session you complete.

Interested? Contact:

participate@taglab.ca

Appendix B: Informed Consent Form

Informed Consent

Please read the following information before beginning the study.

Principal Investigator
Velian Pandeliev, M.Sc. Candidate, 647-864-1472
Technologies for Aging Gracefully Lab, University of Toronto, Department of Computer Science

Faculty Supervisor
Ronald Baecker, Ph.D., Professor, 416-978-6983
Technologies for Aging Gracefully Lab, University of Toronto, Department of Computer Science

Project Purpose and Procedures
The purpose of this project is to evaluate the effectiveness of mental fitness interventions and board games on cognitive performance in older adults. We are also interested in how the online method of delivery affects participant retention, motivation, and whether data collected online can serve to validate mental fitness interventions.

To assist in this regard, you are asked to participate in the following online study. It will consist of an initial questionnaire and playing games over 10 45-minute sessions within 30 days. Your cognitive performance will be evaluated using simple online tasks before and after the experiment, as well as 15 days afterwards. Finally, a semi-structured interview will be completed over a video messaging platform to gather your thoughts and feedback on the study and on the online system. We will maintain contact with you over e-mail. No other identifying information will be collected, your name or e-mail will not be associated with your data, and no personally identifying information will be published.

We expect the study to take approximately **15 hours** to complete over the course of **45 days**.

Potential Risks
There are no risks believed to be associated with this project as participation is voluntary and no individuals will be identified.

Potential Benefits
Information gathered in this study will contribute to understanding how to evaluate mental fitness interventions and how participants respond to online experimentation. This data can make it easier for researchers to conduct studies to validate mental fitness tasks and provide users with better, more reliable information about the benefits they can expect. Participants may also experience minor improvements in executive function, visuo-spatial reasoning and vocabulary.

Remuneration/Compensation
We are very grateful for your participation. While you will receive no compensation for participating, we will donate 3$ to a registered Canadian charity of your choice for every game-playing session you successfully complete.**Confidentiality**
No individuals are identified in this study and identifying personal information will be stored separately from data. Documents and electronic data will be securely kept for five years and then destroyed. Identifying details will be destroyed as soon as the study is complete.

Contact information about the study
If you have questions or desire further information with respect to the study, you may contact Velian Pandeliev at vpandeli@cs.utoronto.ca or 647-864-1472.

Contact information about the rights of research subjects
If you have any concerns about your treatment or rights as a research subject, you may contact the Office of Research Ethics at ethics.review@utoronto.ca or 416-946-3273.

I,____, agree to participate in the study as outlined above.

My participation in this study is voluntary and I understand that I may withdraw at any time.

☐ I agree to my name and contact information being kept on file and I am willing to be contacted very occasionally to see if I would participate in a future TAGLab study. I understand that my name will not be associated with my data and that this puts me under no obligation to participate in future studies.

☐ I request not to be notified of any future TAGLab studies. Please do not keep my name and contact information on file.

Subject Signature: ________________________________ Date _________________________

Printed Name of the Subject:
____

Appendix C: Demographics & Technology Questionnaire

In this questionnaire, you will provide basic information about yourself, including occupation, social and gaming habits and familiarity with the Internet.
As the first step in the study, you will be asked to provide some details about yourself, your occupation, your social and gaming habits and your use of technology. Please don't provide any identifying information in the text fields of this questionnaire.

There are 22 questions in this survey

Background and Demographics

Please tell us a bit about yourself.
The first set of questions asks about your background.
1 In what year were you born? *
Please write your answer here:

2 Please indicate your gender. *
Please choose **only one** of the following:

- ○Female
- ○Male

3 Please indicate your marital status.
Please choose **only one** of the following:

- ○Single
- ○Common-Law
- ○Married
- ○Widowed
- ○Separated
- ○Divorced
- ○Other

4 Compared with other people your age, how would you describe your health? *
Please choose **only one** of the following:

- ○Excellent
- ○Good
- ○Average
- ○Fair
- ○Poor

5 Please indicate the highest level of education you obtained. Use the comment box to specify the degree/program and other relevant details. *
Please choose **only one** of the following:

- ○Less than high school

- High school diploma
- Some post-secondary (college/university)
- 4+ years of post-secondary (college/university)

Make a comment on your choice here:

6 What is or was your primary job or profession? *

Please write your answer here:

7 Do you work full time or part time, or are you retired?

Please choose **only one** of the following:

- Full time
- Part time
- Retired
- Other

8 Where do you live?

Please write your answer here:

Please indicate the country and (optionally) the city or region in which you live. Do NOT provide your exact address.

Hobbies and Pastimes

In this section, we ask about how you spend your free time.

9 How often do you engage in the following physical activities?

Please choose the appropriate response for each item:

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
Team Sports (soccer, hockey, volleyball, etc.)	○	○	○	○	○	○
Individual Sports (tennis, golf, squash, etc.)	○	○	○	○	○	○
Martial arts (karate, muay thai, boxing, etc.)	○	○	○	○	○	○
Outdoor Activities (skiing,	○	○	○	○	○	○

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
snowboarding, jogging, hiking, etc.)						
Working out (free weights training, fitness classes, etc.)	○	○	○	○	○	○
Water sports (swimming, aquafit, etc.)	○	○	○	○	○	○
Cycling	○	○	○	○	○	○
High-intensity Activities (spinning, sprinting, marathon running, etc.)	○	○	○	○	○	○

10 Did we miss any physical activities? Here you can tell us about other things you do that involve physical activity if you wish.
Please write your answer here:

11 How often do you engage in the following entertainment activities?
Please choose the appropriate response for each item:

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
Reading (books, newspapers, etc.)	○	○	○	○	○	○
Films and Television (movies, TV series, documentary, news, etc.)	○	○	○	○	○	○
Listening to music	○	○	○	○	○	○

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
Playing card games (bridge, eucre, poker, etc.)	○	○	○	○	○	○
Playing board games (chess, Monopoly, Scrabble, etc.)	○	○	○	○	○	○
Jigsaw puzzles	○	○	○	○	○	○
Paper puzzles (crosswords, SuDoKu, etc.)	○	○	○	○	○	○
Playing video or computer games	○	○	○	○	○	○

12 Did we miss any entertainment activities? Please enter anything else you do for fun.
Please write your answer here:

13 How often do you engage in the following creative activities?
Please choose the appropriate response for each item:

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
Writing (poetry/prose/journals)	○	○	○	○	○	○
Composing music, playing an instrument or singing	○	○	○	○	○	○
Drawing/sculpting	○	○	○	○	○	○
Dancing (classes/social events)	○	○	○	○	○	○
Drama (acting/directing)	○	○	○	○	○	○

	More than once a day	Once or more a week	Once or more a month	Once or more a year	Not anymore	Never
Sewing/knitting/embroidery	○	○	○	○	○	○
Woodworking/crafts	○	○	○	○	○	○

14 Have we missed any creative activities? Please enter anything else you do on a regular basis as a creative outlet.
Please write your answer here:

Computer Usage

In this last section of the study, we would like to know about your computer usage and the way you use computers in your daily life.

15 When did you first use a computer?
Please choose **only one** of the following:

- ○pre-1960
- ○1960-1969
- ○1970-1979
- ○1980-1989
- ○1990-1999
- ○2000-present

16 What kinds of computers have you used?
Please choose **all** that apply:

- ☐PC (DOS/Windows)
- ☐PC (Lunix/UNIX)
- ☐Mac/Apple
- ☐Mainframe
- ☐Laptop/Notebook
- ☐Tablet PC
- ☐Smartphone (iPhone, Andriod)
- ☐Handheld (PDA/Palm Pilot)
- ☐iPad
- ☐Other:

17 Do you use a computer for work?
Please choose **only one** of the following:

- ○Yes
- ○No

18 Approximately how many hours a week do you spend using a computer for work?

Only answer this question if the following conditions are met:
° Answer was 'Yes' at question '17 [COM03]' (Do you use a computer for work?)
Please write your answer here:

19 Do you use a computer for leisure or personal tasks?
Please choose **only one** of the following:

- ○Yes
- ○No

20 How many hours per week do you spend using your computer for leisure or personal tasks?
Only answer this question if the following conditions are met:
° Answer was 'Yes' at question '19 [COM05]' (Do you use a computer for leisure or personal tasks?)
Please write your answer here:

21 How familiar are you with the following types of computer applications?
Please choose the appropriate response for each item:

	Entirely Unfamiliar	Mostly Unfamiliar	Mostly Familiar	Very Familiar
Word Processor (e.g., Word, Pages)	○	○	○	○
Email (e.g., Outlook, Gmail, Hotmail)	○	○	○	○
Web Browser (e.g., Internet Explorer, Chrome, Firefox, Safari)	○	○	○	○
Spreadsheet (e.g., Excel, Numbers)	○	○	○	○
Graphics (e.g., Photoshop, iPhoto, Lightroom)	○	○	○	○
Presentation Software (e.g., PowerPoint, Keynote)	○	○	○	○

	Entirely Unfamiliar	Mostly Unfamiliar	Mostly Familiar	Very Familiar
Database (e.g., MySQL, Oracle)	○	○	○	○
Music/Video (e.g., iTunes, Quicktime, Winamp)	○	○	○	○
Computer Games	○	○	○	○
Social Media (e.g., Facebook, Twitter, blogging)	○	○	○	○
Communication (e.g., MSN, ICQ, Skype)	○	○	○	○

22 Which of the following have you done with a computer?
Please choose **all** that apply:

- ☐ I have customized options or preferences within a computer application.
- ☐ I have made a purchase online.
- ☐ I have installed a computer application.
- ☐ I have installed an operating system.
- ☐ I have formatted a hard drive.
- ☐ I have attached a new external device (i.e. printer, scanner, camera)
- ☐ I have connected to a wireless network.
- ☐ I have attached a new internal device (i.e. hard drive, memory, graphics card)

Thank you for completing the survey! Your responses have been recorded. You will receive an e-mail shortly confirming this for your records.

Appendix D: Selective Reminding Task

IMMEDIATE RECALL

Buschke, H. & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. Neurology, 24, 1019-1025.

Introduction:
The Selective Reminding Test (SRT) is a test of verbal learning and memory. The test consists of three parts: Immediate Recall (consisting of 6 trials and represented by columns 1-6 on the score sheet); Delayed Recall; and Delayed Recognition. Delayed Recall and Recognition are administered 15 minutes after termination of Immediate Recall. The subject is not forewarned that the Delayed memory tasks will be administered.
To administer the SRT, you need a score sheet, stopwatch, and the appropriate stimulus cards with the multiple-choice items for the delayed recognition trial.

Instructions:
You may repeat all or part of the instructions if necessary.

"This is a memory test. I am going to read you a list of words and ask you to remember as many of them as you can. First, I will read the words, and then you will tell me the words you remember. Then I will remind you of the words that you did not say, and I will ask you to tell me the whole list again; that is, the words you said the last time and the ones I reminded you of. We will do this several times. Each time you will try to tell me the entire list, and I will remind you of the words you left out. Do you understand?"

The first time I read the list I would like you to repeat each word aloud to be sure I said them clearly enough. Ready?" (if necessary, you may add, "Now try to remember the *words as you repeat after me.")*

Read the words and then say, **"Now tell me all the words you can remember."** Start timing immediately after reading the word list and telling the subject to begin recall.

To record the words recalled, write numbers indicating the order in which they were recalled in the appropriate column for each trial. (Words do not have to be recalled in the order in which they were read.) After 60 seconds, say: **"Now, I will remind you of the words you did not say and then ask you to try to recall the whole list."** Read only the words that were not recalled on the previous trial at a rate of approximately 1 per second. Again record the responses in the order they are given. Repeat this procedure for trials 2-6.
Testing of Immediate Recall is terminated when the subject recalls all 12 words on two consecutive trials, or when all 6 trials have been completed, whichever occurs first. When Immediate Recall is terminated, record the time on the testing sheet and determine the time when Delayed Recall must be initiated by adding 15 minutes to the present time.
In the interval between immediate and delayed recall, administer nonverbal tasks (e.g., Identities & Oddities, Color Trails, Rosen Drawing, etc.).

Special situations:
Sometimes subjects wish to repeat words read to them in trials 2-6. While this practice should be discouraged, it can be permitted, especially if the subject has had trouble hearing the words. If this happens, the tester may say: **"Just listen and try to remember the words."**
Frequently, particularly on the second trial, the subject will give only the words s/he was reminded of (i.e., s/he forgot on the first trial). If this happens, the subject can be reminded: **"Try to give me the words from the last time that I did not remind you of."**
If the subject recalls a word 3 or more trials in a row and then does not recall it on the following trial, the tester may say: **"There is a word that you have given me quite a few times but you have not said it yet this time."**
The subject can also be told, **"You can run through the list out loud to make sure that you have not left anything out."** This is the maximal amount of prompting that the subject is allowed. Do not spell or define words for the subject.

Intrusions:
An intrusion is a word that is not on the list. The first time that any word that is not on the list is reported, record it in the intrusions section in the column appropriate for that trial, and tell the subject, **"That word is not on the list."**
If the subject reports this word again on any later trial, record it without saying anything to the subject.
The testing is terminated when the subject recalls all 12 words on two consecutive trials, or when all 6 trials have been completed if this criterion is not achieved. If subject recalls all 12 words on any one trial, stop/reset timer, and say: **"Now, try to give me the list again."** Start timing.
When testing is terminated, record the time of day on the testing sheet and determine the time when delayed recall must be initiated by adding 15 minutes to the present time.

SRT Scoring:
Total Recall: Count the number of words reported in each trial/column and record this total in the line labeled total recall. Maximum Score: 72. If subject recalls all 12 words on two consecutive trials before all six trials are administered, all remaining trials receive a score of 12.

Long-term Storage (LTS): If a subject correctly recalls an item on two consecutive trials, that item is assumed to have entered long-term storage. To compute LTS, for each word, find the first instance of recall on two successive trials. That is,the first time that a word isrecalled on two trials without an intervening reminder. Draw a line through the output numbers for that word on those two trials and extend this line to the end of the task (i.e., through the column for trial 6). The first instance of recall on two successive trials in the absence of reminders is the earliest indication that an item has been entered into long-term storage (LTS). From this point on, when this item is recalled it is said to have been retrieved from LTS. Find the number of words in long-term storage on each trial by counting the number of squares (with or without retrievals) crossed by a line on each trial. Record this total on the line next to LTS. Maximum Score:72

Long-term Recall (LTR): Long-term recall reflects the frequency with which words in long-term storage were actually recalled. To compute LTR, find the number of items recalled from LTS on each trial by counting the number of cells with retrievals crossed by a line on each trial (i.e. those indicated by both a number and a line). Record the total words retrieved from long-term storage on each trial in the appropriate cell next to the line labeled LTR. Maximum Score: 72

Consistent Long-term Retrieval (CLTR): Items in LTS may or may not be perfectly recalled. Consequently, we may be concerned with the consistency of retrieval from long-term memory. CLTR begins at the point where the subject consistently recalls an item without error for the remainder of the test. The item must be recalled correctly on at least trials 5 and 6 to be scored under CLTR. Maximum Score: 72

Intrusions: The first time that the subject provides a word that was not on the list, the word is counted as a DIFFERENT INTRUSION. Subsequently, if the subject repeats an intrusion, the repetition is tallied under TOTAL INTRUSIONS. NB: DIFFERENT INTRUSIONS are not also tallied under TOTAL INTRUSIONS. Only repetitions of intrusions are tallied under TOTAL INTRUSIONS.

DELAYED RECALL & RECOGNITION

Delayed Recall:
15 minutes after termination of Immediate Recall, the participant is asked to recall and then recognize the words that were on the list. The participant must have no indication that this testing will occur. At the appropriate time, the participant is told the following: **"Remember the list of words that you were trying to learn a little while ago? You were trying to recall the whole list of words and I was reminding you of the ones you forgot. Now I would like you to try to recall as many of the words on this list as you can."** Start timing immediately after completing these instructions. Allow 1 min for delayed recall. Mark the words in the order that the participant recalls them in the delayed recall column.

Delayed Recognition:
If all 12 words are not recalled, a recognition trial is given only for each word that was not recalled. The multiple-choice items to the right of the delayed recognition column are typed on pages, one page for each 4-word group. Present the page for the first word that the participant omitted on delayed recall with the following instructions: **"One of the words on this card was also on the list of words that you had tried to learn. The other three words were not on that list. Can you show me which word was on the list?"** Point to each word and say it aloud while the participant is looking at it. If the participant is unsure which word was on the list, encourage guessing and say, **"One of these was on the list, can you take a guess?"** Continue the delayed recognition testing for all words that were omitted on the delayed recall trial. There is no time limitation for delayed recognition testing.

Scoring:
Delayed Recall: The delayed recall score is simply the number of words recalled during the delayed recall trial. Maximum Score:12

Delayed Recognition: The delayed recognition score is the number of words correctly recognized during delayed recognition testing PLUS the number of words correctly recalled during delayed recall. Maximum Score: 12

STUDY ID ____

DATE ________

SRT(FORM 3) DELAYED RECALL

	RECALL	RECOGNITION		
1. THROW	__	__	THROW THROUGH	TOSS PLATE
2. LILY	__	__	FLOWER INTENT	LILT LILY
3. FILM	__	__	FILM SLAVE	MOVIE KILN
4. DISCREET	__	__	WAVER DISCREET	CAUTIOUS DISTRICT
5. LOFT	__	__	SOFT ATTIC	LOFT TACK
6. BEEF	__	__	BEET CLUE	MEAT BEEF
7. STREET	__	__	STREAM SPEED	STREET ROAD
8. HELMET	__	__	HELMET BACON	ARMOR VELVET
9. SNAKE	__	__	SMOKE SNAKE	SERPENT POOL
10. DUG	__	__	HOED HAY	DUG DOG
11. PACK	__	__	BLANK PACK	BUNDLE PUCK
12. TIN	__	__	TON FOIL	SHIRT TIN
TOTALS	_____	_____		

SELECTIVE REMINDING TEST (FORM 3) Date______ Examiner________ ID______

	1	2	3	4	5	6
THROW						
LILY						
FILM						
DISCREET						
LOFT						
BEEF						
STREET						
HELMET						
SNAKE						
DUG						
PACK						
TIN						

remind words ... words she ...
Q: and any o... you said b...

	1	2	3	4	5	6
INTRUSIONS						

TOTAL INTRUSIONS__

DIFFERENT INTRUSIONS_

	1	2	3	4	5	6	TOTALS
TOTAL REC							
LT RECALL							
LTS							
CLTR							

Appendix E: Boggle Scoring Rules

Word Length	Points Value
Not in dictionary	0
<3 letters	0
3 letters	1
4 letters	1
5 letters	2
6 letters	3
7 letters	5
8+ letters	11

Appendix F: Debriefing Interview

This semi-structured interview will be used to get a subjective idea of how participants felt about the portal. Questions to be answered include:

- Is it okay if I record our conversation so I can refer to it later?
- should take 20-30 minutes

How easy was it to learn to use the portal?

Would you consider using the portal for another study?

If you were able to keep using the portal to play games, would you?

Were you satisfied with the level of personal contact with the experimenter?

How does online experimentation compare to going into a lab?

What motivated you to keep going with the study?

Are there other rewards you would have preferred besides donating to a charity in your name?

Did you enjoy playing games on the portal?

What would you like to see improved in the portal?

How did you decide when to log on and complete sessions?

How do you feel about playing online games?

Would you be interested in playing more online games outside of a study context?

Could you e-mail me the name of the charity to which you would like to donate the money you earned by participating in the study?