The Rise of the Super Experiment

John C. Stamper
Carnegie Mellon University
Pittsburgh, PA
john@stamper.org

Derek Lomas
Carnegie Mellon University
Pittsburgh, PA
dereklomas@gmail.com

Dixie Ching
New York University
New York, NY
dixie@nyu.edu

Steve Ritter
Carnegie Learning, Inc.
Pittsburgh, PA
sritter@carnegielearning.com

Kenneth R. Koedinger
Carnegie Mellon University
Pittsburgh, PA
koedinger@cmu.edu

Jonathan Steinhart
Vienna, Austria
jsteinhart@gmail.com

ABSTRACT

Traditional experimental paradigms have focused on executing experiments in a lab setting and eventually moving successful findings to larger experiments in the field. However, data from field experiments can also be used to inform new lab experiments. Now, with the advent of large student populations using internet-based learning software, online experiments can serve as a third setting for experimental data collection. In this paper, we introduce the Super Experiment Framework (SEF), which describes how internet-scale experiments can inform and be informed by classroom and lab experiments. We apply the framework to a research project implementing learning games for mathematics that is collecting hundreds of thousands of data trials weekly. We show that the framework allows findings from the lab-scale, classroom-scale and internet-scale experiments to inform each other in a rapid complementary feedback loop.

Keywords

eScience, experiment design, internet scale

1. INTRODUCTION

Web-based software is creating an explosive growth in the use of randomized controlled experiments in education, due to the relative ease with which users can be randomly assigned to different experimental conditions. Scientists are beginning to recognize the coming data surge and developing new ways of analyzing data at "internet scale." The vastly increased scale of subject populations online can produce a categorically different mode of experimentation in education. For this reason, we propose a new experimental framework that takes advantage of rapid internet-scale experimentation, while retaining the control of lab-scale and classroom-scale experiments.

Randomized controlled trials are regularly used to drive design decisions on the internet. In its simplest form, A/B testing is a form of experimentation where one of two advertisements are randomly delivered to each incoming site visitor. This allows advertisers to determine which advertisement results in improved outcomes (such as a greater click-through rate) [3]. Multiple tools exist to support website optimization, including the free Google Site Optimizer that supports both A/B tests and multi-variable testing. Recently, free-to-play online game companies, such as Zynga, have made use of large-scale optimization experiments with their large number of online players. By randomly assigning players to hundreds of different game design configurations, they can optimize the game design to maximize the conversion of players to paying customers [7].

2. Internet Scale Research in Education

Internet-scale research introduces new potential methods in Educational Research. For instance, optimization experiments like Response Surface Methods, are a common applied research method for improving industrial process outcomes. These experimental designs showed early promise for improving educational outcomes [5], but because the designs would have required many hundreds of students, they were expensive and impractical. Internet-scale research can now support these optimization experiments, along with these other experimental advantages:

Increased number of conditions. With tens of thousands of "user-subjects," internet-scale research studies present the opportunity for researchers to run dozens—even hundreds—of different experimental conditions simultaneously. This easily contrasts with lab or field-scale studies, where available resources and subject pools typically constrain experimental designs to fewer than 8 experimental conditions. Furthermore, with fewer conditions, experiments can be conducted within days, rather than months.

Ability to measure "true" task engagement. Internet-scale research is also uniquely suited for measuring task engagement. Because the researcher typically lacks control over participants (they can quit far more easily than in lab or classroom experiments), the internet is an ideal setting for investigating user motivation. If players assigned to condition A play significantly longer than players in condition B (i.e., were engaged in the task for longer), then condition A can be said to be more engaging than condition B. The ability to measure and compare engagement makes it possible to measure how different design elements and configurations affect player engagement.

Increase in external validity. A third advantage of internet-scale research is the high external validity—experiments are conducted with actual "real-world" users. While the lack of control over subjects can result in noisy data, this noise is useful for preventing over the over-fitting of predictive models that constructed for use "in the wild."

Greater access to all users. A fourth advantage of internet-scale research is the fact that informed consent is not required if the users are anonymous. Even with educational exemptions to informed consent, parental opt-out forms can still pose a barrier to many field-based educational studies. While researchers could

potentially make use of informed consent (and thus obtain non-anonymous data), anonymous data collection is likely to remain a characteristic of most large internet-scale research.

Of course, the lack of information about participants is also a key drawback of internet-scale research. Broadly speaking, internet scale studies cannot collect rich information about participants. Therefore, these studies are unlikely to be suitable when research questions require demographic data, detailed pre/post tests, participant observation, talk-aloud protocols, or any kind of psychophysiological measure. Finally, the lack of participant control means that internet scale studies may not be appropriate if repeated participation over time is required.

Given these drawbacks, it is clear that traditional lab based experiments and structured field trials still provide valuable data that internet scale experiments cannot. However, there is much to be gained from internet scale studies. The Super Experiment Framework (SEF) seeks to illustrate how different scales of experimentation can productively inform one another. The SEF framework, seen in Figure 1, is split into three general experimental parts that are roughly delineated by scale. Lab-Scale experiments are smaller highly controlled studies that take place in a lab or single classroom, generally not exceeding 50 participants. School-Scale experiments are formal experiments that take place in multiple classrooms or schools consisting of hundreds to thousands of participants. Internet-Scale experiments are informally delivered online to thousands to millions if participants.

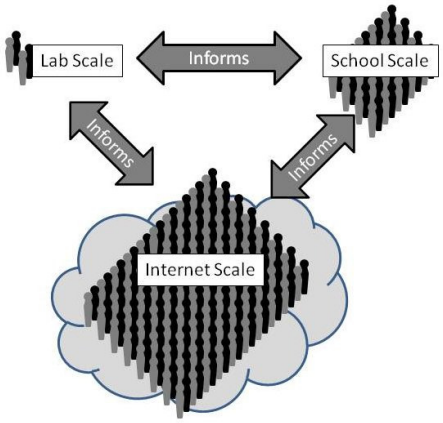

Figure 1. The Super Experiment Framework showing how each of the component scales informs the others.

In the SEF framework, each component provides an experimental level that can be used to answer specific questions that might be difficult or impossible to answer using one of the other components. Further, the various components can be used to expand or validate findings of the other components. A feedback loop can also be used with the framework where internet scale experiments can identify areas of focus for lab scale experiments, which can then be validated in school scale experiments. An overview of each of the SEF components can be seen in Table 1.

School scale and lab scale experiments typically recruit subjects and then randomly assign them to different experimental conditions as part of a single experiment. However, internet-scale research creates situations where multiple experiments are randomly drawing from the same pool of subjects. Just as a single experiment contains multiple experimental conditions, the SEF contains multiple experiments. Because the different experiments are derived from the same pool of random assignment, experimental conditions that are not part of the same experiment may still be compared to one another, if desirable. While there may be few immediate benefits of this comparison, the super experiment is a unique characteristic of internet-scale research. Therefore, the use of the term "super experiment" in the super experiment framework simply refers to the broad network of information flow between different scales of experimentation, from the lab scale, to the school scale and to the internet scale.

Type	Benefits	Drawbacks
Lab Scale (1-75)	Rich user data, Formal, Controlled CTA, Talk alouds, Psycho-physiological studies	Effect Size, Replication, Scalability, Experimenter effects, Threats to external validity
School Scale (25-10,000)	Formal, Controlled, Validation, Good randomization, Surveys, Enforced participation	Expensive, Difficult to replicate, Threats to external validity
Internet Scale (10,000-millions)	Informal, Large data collection, Rapid, High external validity, Decreased Type II error rate, High power	Anonymity, High attrition, Data overload, Threats to internal validity

Table 1. Components of the Super Experiment Framework

3. IMPLEMENTATION EXAMPLE

The need for the SEF framework was initiated through our work in creating online games for learning. The number of potential experiments was large and the opportunity to field the games at each of the scales identified in the SEF framework provided the need to build a feedback loop to execute many experiments at internet scale in order to narrow down the potential experiments to test at the more controlled school scale. "Battleship Numberline" (BSNL), an online educational game, benefits from the super experiment framework.

Designed to improve number sense among elementary and middle school students, BSNL provides practice estimating numbers on a number line within four content domains: whole numbers, fractions, decimals and measurement [4]. The game narrative involves defending Numbaland Island from invading robot pirates by firing projectiles at their ships and submarines. BSNL involves two basic modes: naming numbers and placing numbers. In the naming condition, players type a number that corresponds to the location of an enemy ship that is positioned on a number line between two marked endpoints. In the placement mode, the player is given the numeric location of a hidden submarine (e.g., "Submarine spotted at 1/3") and needs to click on the location that they believe corresponds to the number. After the player has typed a number or clicked on the number line, a projectile drops vertically from the top of the screen to the designated location on the number line. Animation and text-based feedback communicates the player's accuracy after every round.

A primary goal of our research has been to understand how different game design factors affect player learning and engagement. In order to systematically investigate these factors, we implement these design factors as flexible xml-based parameters that can be determined at the game runtime. We are

then able to create online experiments that randomly assign new players to a set of different game sequences.

During gameplay, BSNL generates an online data log of the task context (the above xml parameters) along with data describing the player's performance on each opportunity. On each item, we log the player's reaction time, their accuracy, and a binary field indicating whether the player was successful or not. Logs are then imported into the PSLC Datashop [2], which allows for the secondary analysis of player performance and learning. The hit rate measure is essential for enabling Datashop to plot learning curves of error rate over time. By labeling different items in the game with different knowledge components (e.g., reducible fractions, unit fractions, etc), we can plot learning curves for each knowledge component. Learning curves can also be described based on fluency [1], where we plot the reduction of reaction time over opportunities played. In addition to these measures of learning and performance, we investigate player engagement through two measures: the total number of items played and the total amount of time spent playing. These two metrics correspond with our construct of intrinsic motivation or player engagement.

The number of potential parameter settings in BSNL makes it a great tool to answer many research questions, but at the same time the number of possible settings make it difficult to decide on what settings to in traditional lab or school settings. For this reason, it is a perfect candidate for use in the SEF. Next, we show how the results of different types of experiments at one scale inform new experiments on a different scale.

Lab Scale informing School Scale. The use of a lab experiment to inform a field trial at a school is one of the most common types of experimental design. It is still an important part of the SEF. We performed a lab scale experiment, which is now being validated at the school scale. This experiment was conducted at a small Catholic liberal arts University. Although the college is co-educational, its focus is on women's education, and 89% of the participants were women. Participants were 18 students in an eight-week first-year seminar course, which met once per week. Students chose for this seminar period to focus on mathematics games. Over 5 weeks, we administered a short (typically one minute) paper-and-pencil pretest, asked students to play a specific fluency game for approximately one-half hour and then gave a posttest which was identical in content to the pretest. In all but the first week, the pretest was preceded by a delayed post-test, which was a repeat of the posttest from the previous week's materials.

In four of the five experiments significant improvement was shown on a delayed post-test, and three of the five showed immediate results. Effect sizes were also quite large, ranging from 0.4 to 2.4, indicating that these results are not only significant but substantial. Prior to the first experiment, students were given a survey about their confidence in mathematics (containing questions like "I am sure that I can learn math.") and about text anxiety (containing questions like "I am so nervous during a test that I cannot remember facts that I have learned"). The two scales were mixed in a 16-item form. Students were asked to rate each statement from 1 ("strongly disagree") to 5 ("strongly agree"). Student confidence increased significantly, $t(14)=-3.2$, $p<.01$, $d=0.4$, but there was no change in test anxiety, $t(14)=-3.1$, n.s.

Due to the success of this lab scale experiment, a similar school scale experiment is now being conducted in multiple college classrooms over an entire semester. Unlike the lab scale, the researchers are not present in these classrooms, but we expect to see similar results.

School Scale informing Internet Scale. BSNL was designed based on an existing body of literature that investigated number line estimation in the laboratory [6]. The game was playtested with 8 elementary school students, to refine usability issues in the design. Following this, a school scale study was conducted with 119 students in grades 4-6. Students showed significant improvement in hit rate form the first to second opportunity (see Figure 2), and students demonstrated significant improvements in the estimation of fractions on a number line after 20 minutes of gameplay. Moreover, 82% of players (74% females, 92% males) reported that they wanted to play the game again [4]. The data from these classroom studies was imported into the PSLC Datashop to test various knowledge component (KC) models. We identified a KC model based on the various regions of the number line. This knowledge component model was then used to produce a Bayesian Knowledge Tracing adaptive sequencing algorithm. This algorithm was then tested online in comparison with a randomly sequenced level. Preliminary results suggest that the BKT adaptive sequence did not result in significantly greater player engagement than the random sequence.

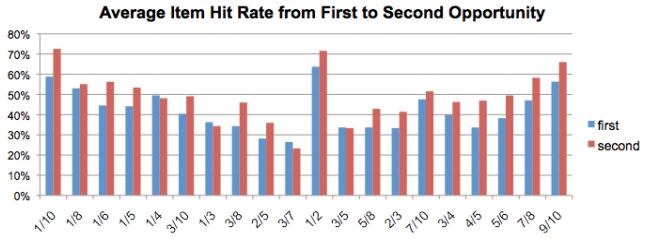

Figure 2. Illustrates the average improvement from the first opportunity to the second opportunity, by item presented. The clear patterns of difficulty are used to generate knowledge component models in Datashop.

Internet Scale informing School Scale or Lab Scale. Internet-scale experiments can be useful for documenting the difficulty of different task configurations. This is useful in the field of EDM, as it allows for the generation of knowledge component models. Different tasks are said to require different knowledge components if and only if the tasks result in different performance rates or learning curves. Therefore, by assessing the difficulty of instances over a broad task design space, we can understand how the task design space maps to various KC models.

For example, Rittle-Johnson, Siegler and Alibali found that tickmarks supported the estimation of decimals on a number line [6]. In order to replicate this work and extend it, we randomly assigned online players to 6 different conditions in both the decimal and whole number domain. Players either encountered tickmarks dividing the number line into tenths, fourths, thirds, halves (midpoint), or no tickmarks at all. Finally, an additional two conditions looked at the interaction of an adaptive sequencing algorithm with tickmarks at the midpoint. An overview of the experiments and conditions can be seen in Table 2. Over 80,000 internet users participated in the experiment.

An experiment with this many conditions would be difficult to replicate in a lab or classroom. This broad investigation of the effects of guides enabled us to observe two unusual outcomes. First, there was an apparent interaction effect between our adaptive sequencing condition (termed "ITS") and the midpoint

guides. Neither Second, the 10th guides apparently increased player engagement in the decimal condition, but decreased engagement in the whole number condition. These insights have led us to execute similar lab scale experiments to replicate and better understand these specific results.

Experiment Name	Conditions	Players
Adaptive Sequencing	15	19,856
Difficulty Sequencing	6	6,302
Difficulty Comparison	6	6,234
Expanded fraction set	4	5,596
Guides Engagement	10	11,386
Guides Learning	20	22,441
Measurement Study	3	10,014
Total	**64**	**81,829**

Table 2. List of experiments running concurrently with a total of 64 conditions.

4. CONCLUSIONS AND FUTURE WORK

Technology is forever changing the way we conduct experiments. The traditional paradigm is no longer the best way to do things. Data is coming in faster, larger, and more fine grained. Instead of focusing eScience efforts in just analyzing we have created a framework to exploit internet scale experiments, while still creating valid findings in real classrooms.

The main contribution of this work is the development of the Super Experiment Framework which incorporates a feedback loop allowing for experiments of different scales to inform each other. This has become possible, and even necessary, with the use of the internet to collect a large amount of experimental data. Internet scale allows for optimization experiments that would be too expensive to do at field level. This is truly applied educational research that, as we have shown, provides insights that can inform more controlled lab or school scale experiments. We also explained our initial implementation of the SEF with a large project with broad scope and many interesting research questions. Traditional "one-way street" experiments of lab to school are slow to findings and outdated. Our work shows how utilizing all three scales of experiments leads to rapid findings that can lead to real implementable insights efficiently.

Making the framework possible is the accessibility of internet scale experiments. The key barrier to internet scale educational research is attracting large numbers of users. Research projects rarely invest in high-quality software design and usability, which is usually necessary to achieve widespread adoption. However, once this quality is developed, large numbers of users can be reached through collaborations with one of many internet portals that seek to aggregate educational content (e.g., Brainpop.com).

Another challenge is instrumenting software for generating data logs that measure player performance, learning and engagement. Log files should capture not only correctness information, but the amount of time that players spend on an activity, as well as the number of opportunities attempted to make these measures.

A third challenge is the configuration of the software to allow for experimental designs. This involves the abstraction of design variables in the software's design space, such that different instances of the software can be created quickly. For instance, we use xml to define game levels at run-time. These configurations can then serve as different experimental conditions that can be randomly deployed to online users.

Finally, one unusual new challenge in internet scale research is the efficiency of subject-pool utilization. While lab or school scale researchers expend significant effort to recruit a sufficient number of subjects in order to achieve statistical significance, internet scale researchers increasingly face the challenge of making use of tens of thousands of subjects in an efficient manner. Certain types of experimentation may result in inconsistent user experiences that reduce overall participation.

Some challenges will be particular to individual experiments. For instance, in our online experiments we observe strong seasonal effects of weekends and school holidays, where the number of players is greatly reduced. This suggests that certain experimental comparisons should be sensitive to the time period of the study, not merely the number of subjects.

Many of these challenges can be mitigated by validating the results of internet scale experiments with controlled classroom experiments. As shown in the experiment section, we are continuing to run a number of experiments of scales based on findings of different scales. This feedback loop will continue in the future as we strive to optimize the games to maximize learning. We believe this framework will rapidly lead to significant discoveries that are replicable at each of the scales.

5. ACKNOWLEDGMENTS

We would like to thank the Pittsburgh Science of Learning Center, the DataShop staff, the Next Generation Learning Challenge, Carlow University, and Pellisippi State University for supporting this research.

6. REFERENCES

[1] Baker, R., Habgood, M., Ainsworth, S., & Corbett, A. Modeling the acquisition of fluent skill in educational action games. User Modeling, 4511, (2007), 17-26.

[2] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2011) A Data Repository for the EDM commuity: The PSLC DataShop. Handbook of Educational Data Mining. CRC Press

[3] Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery, 18(1), (2008), 140-181.

[4] Lomas D., Ching D., Stampfer, E., Sandoval, M., Koedinger, K. Battleship Numberline: A Digital Game for Improving Estimation Accuracy on Fraction Number Lines. *Conference of the American Education Research Association (AERA)* (2012).

[5] Meyer, Donald L. Response Surface Methodology in Education and Psychology, Journal of Experimental Education, 31, 4, (1963), 329-336.

[6] Rittle-Johnson, B., Siegler, R. S., and Alibali, M. W. Developing conceptual understanding and procedural skill in mathematics: An iterative process, *Journal of Educational Psychology*, 93, (2001), 346-362.

[7] Sheffield, B. GDC Canada: Bill Mooney Outlines Zynga's Methodology For Success, Gamasutra, May 6, 2010. Retrieved 2/10/12:http://gamasutra.com/view/news/28434/