NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments

Xiaojun Meng, Pin Foong, Simon Perrault, Shengdong Zhao

NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments

Xiaojun Meng[1], Pin Sym Foong[1], Simon Perrault[2], Shengdong Zhao[1]

[1]NUS-HCI Lab, National University of Singapore
Singapore,117417
{xiaojun, zhaosd}@comp.nus.edu.sg
pinsym@u.nus.edu

[2]Yale-NUS College
Singapore, 138529
simon.perrault@yale-nus.edu.sg

ABSTRACT. In this paper, we introduce NexP (Next Experiment Toolkit), an open-source toolkit for designing and running controlled experiments. Unlike previous toolkits, it is targeted for the unmet needs of the beginners in experimental design, who may not have had prior statistical training, or experience in creating, implementing and executing controlled experiments. To accommodate such users, NexP features a hypothesis development process that scaffolds beginners into bridging the gap between daily language and formal statistical language. In our evaluation, we compared NexP against a state-of-the-art experimental design toolkit. Results showed that novices considered NexP more intuitive and easier to use. Users also reported that NexP helped them to better understand the experimental design process, making it a useful tool for both productivity and education.

Keywords. NexP; Controlled Experiment; Toolkit; Design Platform.

1 Introduction

Controlled experiments are an important, widely-used research method in HCI to evaluate user interfaces, styles of interaction, and to understand cognition in the context of interactions with systems [6, 10]. Even for experienced researchers, designing a controlled experiment can be a tedious, multi-step process that is prone to errors.

To reduce the tedium involved in controlled experiment design, HCI researchers have proposed various toolkits to facilitate the design of controlled experiments [12, 17, 20]. Of these, only Touchstone [17] is a complete and functional general-purpose platform that facilitates exploring alternative designs of controlled experiments. It documents the experimental design in a shareable digital form to support replication and extension of previous research work in the HCI literature.

However, because of the complexity of controlled experimental design, existing toolkits continue to challenge beginner researchers. In our preliminary study with six beginner researchers, we saw that they had difficulty in identifying different factors under a general research question, translating a research question into the appropriate variables and determining an appropriate arrangement of conditions, trials, blocks. It

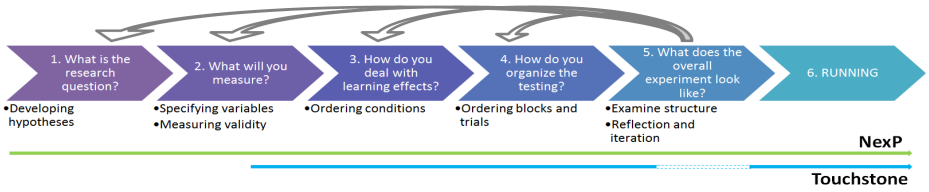

Fig. 1. Conceptual Structure of NexP compared to Touchstone. NexP scaffolds beginner researchers from ideation to implementation and back, while Touchstone mainly provides automation to experienced users in certain steps.

seems that beginners lack the ability to gradually refine questions from a rough inquiry to a robust, testable hypothesis.

There is evidence in pedagogical literature [1, 15] that these may not just be transient novice difficulties, but have an additional layer of conceptual difficulty associated with experimental design and statistical knowledge. In their statistical pedagogy textbook, Ben-Zvi and Garfield [1] identify a need for "selective, thoughtful, and accurate use of technological tools and increasingly more sophisticated software programs" that not only support doing the task, but support learning *about* that task.

Taken together, these beginner shortfalls run counter the usability of existing experimental design tools. Hence, we were motivated to create an experiment design tool that can better facilitate the learning of experimental design and is easier for users who might have a low understanding of statistical terminology and a low working understanding of the purposes and goals of experiment design.

We implemented our approach into a web-based open-source toolkit called NexP that has both design and execution platforms (**Fig. 1**). Compared with previous experimental design tools we introduced 1) a Question-Answer Structure that connects initial experimental concepts of a beginner researcher to the specialized domain of experiment design and 2) a streamlined step-by-step process to guide such users through complex process. With these two features, our goal was to enable users to explicitly link the final study design product with the study design decisions made previously.

To evaluate the effectiveness of this design platform, we conducted a workshop study where experimental design novices compared the usage of NexP and Touchstone [17] in several experimental design tasks. We found that participants preferred NexP because of the guided process of hypothesis formulation. In addition, participants also indicated that NexP enhanced their understanding of the experimental design process. Overall, the contribution of this paper is the following:

1) We propose a 5-stage approach using a Question-Answer Structure to enhance the users' understanding of experimental design by gradually guiding the beginner researchers from an initial idea to a concrete design. We implemented it into a web-based open-source toolkit NexP [8, 9].

2) We evaluated NexP's experiment design process against an alternative toolkit with beginner researchers in terms of ease of use and learning facilitation for experimental design. Our results indicated that NexP is easier to use and better supports the participants' thinking and learning of experiment design.

2 NexP structure

NexP's basic structure is a 5-stage process for designing experiments, with an additional 6th stage for running experiments. Much of the full process is detailed elsewhere in our 2-page implementation demo paper [18] and in our CHI course material [13]. Our current version transformed the previous template approach into the Question-Answer approach to better guide beginner users. It also incorporated a running platform into the previous implementation. This Question-Answer structure approach (elaborated below) is informed by the practice of scaffolding problem solving [7], and case of instructional design for complex domains [2, 11] Interim pilot testing showed that participants found this guidance useful. We adopted this approach for Stages 1-4.

As a result, the designing stages were focused on scaffolding the beginner researchers who may not understand statistical terms, or may not have had sufficient experience with experimental design to understand the purpose of such terminology. The elements of experiment design were deconstructed into 5 stages, and each deals with one section: Hypothesis, Variables and Measurements, Test Conditions, Test Order and Study Logistics.

Question-Answer-Structure Approach

Each of the first four stages is constructed in a similar manner – a series of questions, followed by a summary that restates the answers in more academic and statistical terms, thus bridging the gap from informal, non-scientific thinking, to generating formalized structured hypotheses. We illustrate this process with the transition from Stage 1 to 2.

Stage 1 is an ideation support stage, which helps the researcher in specifying a hypothesis. In this stage we tackle terminology by ignoring it altogether during ideation. Instead we begin by asking the evoking question "What is the general question that you want to answer in the study?" This is followed by unpacking the question further into smaller questions to elicit the comparative claim, the key task(s) involved, the measure used. Each step comes with explanations, help menus and loadable examples (**Fig. 2**).This step closes with a presentation of the answers in a structured 'research claim' statement, populated by the answers to the preceding questions:

In my experiment, I want to hypothesize that for [target user] to do [task], the solution(s) [my idea] is better than [comparable idea]. To test this hypothesis, I will vary the [other contextual factors] and measure [measureable data].

At this point, the user is encouraged to make logical adjustments to their answers if they spot inconsistencies. Also, this is the first introduction to some experimental design terminology.

Stage 2 aims to connect the hypothesis to the variables in a controlled experiment. The terms that refer to the different types of variables, e.g. dependent/independent can be confusing, especially for beginners. Moreover, the specific terminology varies by research field, e.g. independent variable vs. factor. Therefore, in this step we explicitly make these connections by showing the claim statement from the first step as a point of comparison, but also sorting the individual elements into independent and dependent variables in this page, and labeling them as such. In effect the users' gener-

ic language statements are again re-structured into more formal statistical language. The user can freely modify these variables (add, remove, rename, reorder) and add test conditions with descriptions to each independent variable.

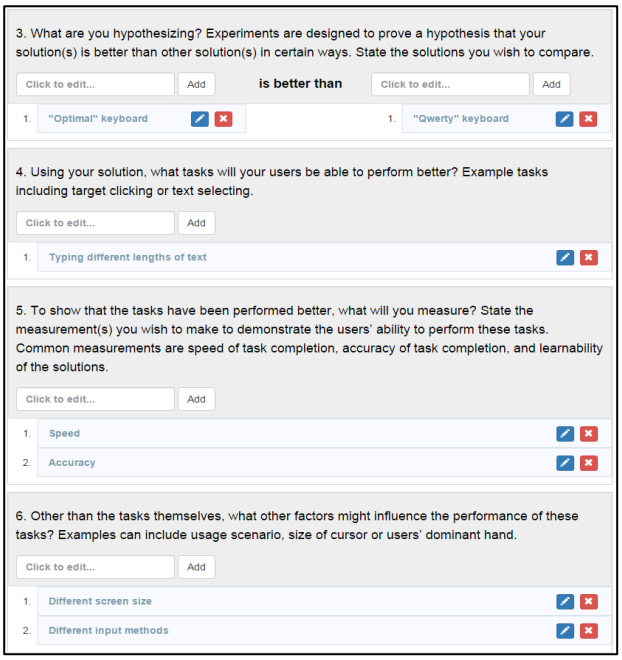

Fig. 2. Illustration of Question-Answer-Structure approach to formalize the hypothesis.

These two stages illustrate how the user is moved from a generic language statement into a claimable, provable claim statement in Question-Answer approach. The remaining stages proceed in a similar, linked manner.

Stage 3 presents both developed statements from Stage 1 and Stage 2, and presents the user with choices about the order of conditions. The terms 'between subjects', 'counter-balancing' for example require higher background knowledge than previous terms. As such, the help text for this stage is much more elaborate, and illustrates both pros and cons of each approach. This is one example of how NexP employs information redistribution to reduce cognitive load.

Stage 4 guides the experimenter to organize the structure of the experiment. It aids in balancing blocks and trials, and working through the timing of the entire experiment. Based on the existing design, NexP suggests the minimum number of participants for the experiment.

At the end of the 4th stage, the user is able to obtain an overview of the entire experiment (**Fig. 3**.a). This overview acts a preliminary simulation of the entire experiment so that the design can be iteratively refined before deployment. In our evaluations, we found that this is the point at which most beginners realize that their experiment is far too elaborate, and they often returned to previous stages to adjust the parameters.

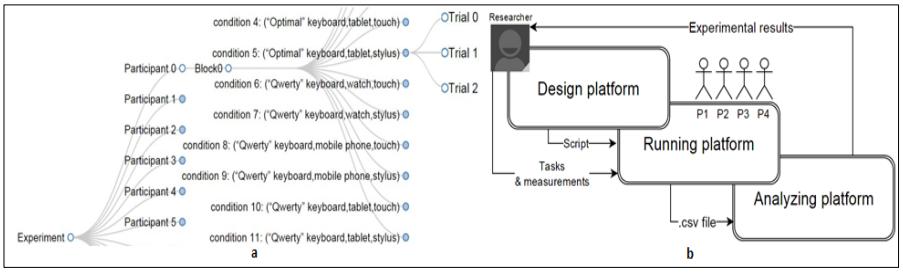

Fig. 3. a) Simulation of the whole arrangement in the experiment; b) Summary of the design, running, and analyzing platform

Stage 4 is often the start of the iterative loop where users move between the stages freely to iterate the parameters and evaluate the results of their adjustments. It is in this iteration process that users experience the elaboration process that conveys the method of developing experimental design. The complexity of the dependent parts that have been distributed among the previous stages becomes clearer and users can freely iterate until the hypothesis, variables, conditions and trials come together in a viable experimental design.

Stage 5 of NexP guides the experimenter to set the instructions and detailed procedures for conducting the experiment, such as providing recruitment forms and consent forms templates. This stage also supports the experimenter in designing the pre-/post-questionnaires using Google Forms and later invokes the designed questionnaires in the running platform. The entire design can then be saved as either a summarized PDF report or a JSON file digitally, which can be imported and shared with other researchers for refinement and re-evaluation in the future.

Executing Experiments

In NexP, we implement a common platform for running different types of controlled experiments leveraging the browser. NexP can automatically read the experimental JSON file and output the whole web-based framework in the arranged order for running the experiment. Experienced programmers can modify the source code of this web framework according to their needs and implement tasks as they require. **Fig. 3**.b summarizes the relationship among design, execution, and experimental analysis stages. The design platform generates a script file that can be executed in the running platform. The measured data (e.g., pre-defined dependent variables) collected by the running platform are saved as CSV files, which can be further analyzed elsewhere.

3 Evaluation

To validate the ease of use and learning facilitation of NexP for people with limited knowledge of experimental design, we conducted a one-day workshop study to compare NexP with Touchstone [17].

Participants and Apparatus: 24 participants (8 female, 16 male) ranging from 21-50 years old (M=29.2, SD=6.9) attended this workshop. The participants included 3 faculty members, 5 researchers and 12 students from several universities, and 4 UX

practitioners. All participants self-reported that they are not familiar with controlled experiments in HCI. None had any prior knowledge of either Touchstone or NexP. Participants were asked to bring their own laptop computers to the workshop, with installation instructions provided earlier.

1. Introduction (60 min)	Introducing experiment design (material from Scott MacKenzie's empirical evaluation book [3])
2. Prep & Training (80 min)	Introducing and installing toolkits; Showing video tutorials and teaching on designing experiments using each toolkit
Lunch break (40 min)	Lunch provided
3. Replication Task (100 min)	Participants replicate two existing controlled experiments, one for each toolkit.
4. Design Task (70 min)	Participants design their own experiment based on a high level description, using a toolkit of their choice.
5. Debriefing (10 min)	Explain the purpose of the workshop; post-evaluation questionnaire

Table 1. Design of the workshop.

Design and Procedure: The entire workshop lasted for 6 hours from 10 am to 4 pm, spread over five sessions as described in **Table 1**. A within-subject design was used. Each participant replicated two tasks using both toolkits (NexP, Touchstone). The order of toolkit use was counter-balanced while the order of the replication tasks remained the same.

In session 3, each participant received soft copies of the original publications [16, 19] and a brief introduction on the replication tasks. Participants were randomly paired into groups of two to complete the tasks. We provided no additional help during this session. Each pair had to submit their resulting files generated by both toolkits to experimenters at the end of the session 3. For each replication task, participants were allowed to spend 35 minutes for a total of 70 minutes.

In session 4, the same participant pairs formed in session 3 were asked to perform the design task by using a preferred toolkit of their choice (either NexP or Touchstone). Finally, they were asked to each fill in the (7-point Likert scale) post-questionnaire based on Computer System Usability Questionnaire [5].

During the debriefing, we explained the complete purpose of the workshop and revealed that we were the authors of NexP. We finally asked participants to sketch features and suggest improvements on paper for an ideal experimental toolkit.

Tasks: *Training Tasks*: The purpose of the training task is to familiarize the participants with the two toolkits. To avoid bias, we created two video tutorials, one for each toolkit, demonstrating the step-by-step procedure to design an interaction technique experiment (Pie Menu vs. Linear Menu from [4]).

Replication Tasks: To minimize transfer of learning effect, participants were asked to replicate two controlled experiments from two CHI publications [16, 19], one for each toolkit. The first study investigated the noticeability of notifications under the influence of feedback modality and physical activity for interactive rings. The second study evaluated performance of different mode switching techniques for pen-based

interfaces. We randomly chose those two studies as they provided clear and appropriate descriptions for HCI beginners to replicate.

Design task: The purpose of the design task was to test how well the toolkits supported participants to design an open-ended controlled experiment. This design task was to compare ShapeWriter [14] to soft keyboards in terms of user performances.

Results

Replication Task: The 24 participants formed 12 groups to work on the replication tasks. 9 pairs (75%) were able to deliver a complete and correct experimental design using NexP, and 4 pairs (33.3%) were able to do so using TouchStone. Of the 3 groups that failed while using NexP, one failed to properly counterbalance the independent variables, another pair did not set the correct number of blocks and trials, and the last pair failed to define the correct independent variables with their levels. For the 8 groups that failed while using Touchstone, three were not able to define the dependent variables. One failed to define the correct independent variables with their levels. Two failed to define the correct counterbalancing strategy. Two were not able to finish on time and thus did not submit the TouchStone design files.

Design task: When the participants were asked to choose their preferred toolkit for this task, 11 out of 12 pairs (91.7%) chose to use NexP. We did not measure the accuracy in this open task, because the detailed design was decided by the individual pairs.

Post-questionnaire: We compared quantitative feedback given on both toolkits using Wilcoxon signed-rank tests. For every affirmation of the questionnaire, the feedback was significantly more positive for NexP (all $p<.05$) than for TouchStone, except on the "It does everything I expect it to do" where NexP performs slightly, yet not significantly better than TouchStone ($M=4.58$ vs. $M=4.25$, $p=.17$).

Questions	NexP	TouchS	p
1. I became more productive with it	5.33	4.67	*
2. Easy to use	4.75	3.21	**
3. It does everything I expect it to do	4.58	4.25	
4. I can use it without written instructions	4	2.54	**
5. It requires the fewest steps to complete task	4.63	3.5	**
6. Using it is effortless	4.25	3	**
7. I quickly became skillful with it	5	3.83	**
8. It is easy to identify the hypothesis	5.29	3.21	**
9. It is easy to identify the independent variables	5.17	3.88	**
10. It is easy to arrange the order of the experimental conditions	5.25	4.25	*
11. I am satisfied with it	5	3.67	**
12. I would use it	5.29	3.63	**

Table 2. Summary of the post-workshop questionnaire. The values in second and third column are the median scores on a 7-point Likert scale for each toolkit. * and ** denote significance ($p<.05$ and $p<.01$).

The results of subjective preferences from the participants (**Table 2**), their choices of toolkit to use in session 4 and the success rate of replicating existing experiments in

session 3, suggest that the scaffolding method embodied in NexP made it easier for non-expert users to use. It also indicates that despite their novice status, they are more capable of designing and presenting complete experimental designs using NexP.

4 Discussion & Implications

The workshop showed that NexP is a more beginner-friendly tool. Participants of the workshop found NexP easier to use and understand because it "*presents information more clearly*" (P3) and the "*terminology for NexP is easy to understand*" (P8).

We were particularly surprised when some participants mentioned that NexP has fewer steps as compared with Touchstone, because both toolkits actually have a similar number of steps for designing an experiment. Possibly, the perceived difference lies in the contrasting ways complex information is presented to users. NexP's use of scaffolding and information redistribution helped to break the experiment design process down to smaller sets of decisions. However, Touchstone relies on a different approach and aims at being as configurable as possible in order to fit the needs of expert users.

Participants also found that NexP provides better support for the thinking process of experiment design. Four participants particularly highlighted how NexP helped them identify and differentiate Independent and Dependent variables through the process of formulating the hypothesis. Participants (P6, P14) were impressed by "*the ability of NexP to convert the answers to [their] questions to independent and dependent variables"*. P12 stated that he liked "*the hypothesis feature of NexP, as it helps to break down the components of the research topic and quickly focus on the important aspects*". Finally, participants P12 and P24 liked the "*tree structure simulation of NexP*". The qualitative comments of the participants affirm the quantitative findings.

We conclude that of the major advantages of NexP is using the Question-Answer structure as well as providing a systematic and streamlined 5-stage approach to help users to think from a broad question to the final concrete design.

5 Conclusion & Future work

In this work, we enhanced the original NexP with the Question-Answer structure to address the needs of beginner users. We evaluated this approach and found that NexP can help beginners to understand and execute controlled experiments better than a leading alternative. As observed by some our participants (P7, 17, 20, 24), NexP holds promise as a companion pedagogical tool to help instructors teaching controlled experiments. In addition, NexP provides the designing and running platform, as well as data support for analysis, making it one of the first attempts for an end-to-end solution. For future work, we would like to see NexP expanded as a learning support tool by conducting studies to further understand the needs of researchers who want to improve in their experimental design skills. We will also improve and publish the open-source NexP's design and running platform for the use of the community.

6 References

1. Dani Ben-Zvi and Joan Garfield. The Challenge of Developing Statistical Literacy, Reasoning and Thinking. Springer Netherlands, Dordrecht. Retrieved September 1, 2016 from http://link.springer.com/10.1007/1-4020-2278-6
2. David H. Jonassen. 1997. Instructional design models for well-structured and III-structured problem-solving learning outcomes. Educational Technology Research and Development 45, 1: 65–94. http://doi.org/10.1007/BF02299613
3. I Scott MacKenzie. 2012. Human-computer interaction: An empirical research perspective. Newnes.
4. Jack Callahan, Don Hopkins, Mark Weiser, and Ben Shneiderman. 1988. An Empirical Comparison of Pie vs. Linear Menus. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 95–100. http://doi.org/10.1145/57167.57182
5. James R. Lewis. 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1: 57–78. http://doi.org/10.1080/10447319509526110
6. Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research methods in human-computer interaction*. John Wiley & Sons.
7. Mary Beth Rosson and John M Carroll. 1996. Scaffolded examples for learning object-oriented design. *Communications of the ACM* 39, 4: 46–47. DOI=http://dx.doi.org/10.1145/227210.227223
8. NexP online access: www.nexp.site
9. NexP's design platform source code: https://github.com/mengxj08/webnexp NexP's running platform source code: https://github.com/mengxj08/platformframework
10. Paul Cairns and Anna L Cox. 2008. *Research methods for human-computer interaction*. Cambridge University Press New York, NY, USA.
11. Paul Cobb and Kay McClain. 2004. Principles of instructional design for supporting the development of students' statistical reasoning. In The challenge of developing statistical literacy, reasoning and thinking. Springer, 375–395. Retrieved September 1, 2016 from http://link.springer.com/content/pdf/10.1007/1-4020-2278-6_16.pdf
12. R. William Soukoreff and I. Scott MacKenzie. 1995. Generalized Fitts' law model builder. *Conference companion on Human factors in computing systems - CHI '95*, 113–114. http://doi.org/10.1145/223355.223456
13. Shengdong Zhao, Xiaojun Meng, Pin Sym Foong, and Simon Perrault. 2016. A Dummy's Guide to your Next EXPeriment: Experimental Design and Analysis Made Easy. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16). ACM, New York, NY, USA, 1016-1019. DOI: https://doi.org/10.1145/2851581.2856675
14. Shumin Zhai, Per Ola Kristensson, Pengjun Gong, Michael Greiner, Shilei Allen Peng, Liang Mico Liu, and Anthony Dunnigan. 2009. ShapeWriter on the iPhone - from the laboratory to the real world. CHI'09 Extended Abstracts on Human Factors in Computing Systems, 2667–2670. http://doi.org/10.1145/1520340.1520380
15. Soofia Malik. 2015. Undergraduates' Statistics Anxiety: A Phenomenological Study. *The Qualitative Report* 20, 2: 120–133.
16. Thijs Roumen, Simon T Perrault, and Shengdong Zhao. 2015. NotiRing: A Comparative Study of Notification Channels for Wearable Interactive Rings. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2497–2500. http://doi.acm.org/10.1145/2702123.2702350

17. Wendy E. Mackay, Caroline Appert, Michel Beaudouin-Lafon, et al. 2007. Touchstone: exploratory design of experiments. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, 1425. http://doi.org/10.1145/1240624.1240840
18. Xiaojun Meng, Pin Sym Foong, Simon Perrault, and Shengdong Zhao. 2016. 5-Step Approach to Designing Controlled Experiments. In Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16), Paolo Buono, Rosa Lanzilotti, and Maristella Matera (Eds.). ACM, New York, NY, USA, 358-359. DOI: https://doi.org/10.1145/2909132.2926086
19. Yang Li, Ken Hinckley, Zhiwei Guan, and James a Landay. 2005. Experimental Analysis of Mode Switching Techniques in Pen-based User Interfaces. CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, 461–470. http://doi.acm.org/10.1145/1054972.1055036
20. Yves Guiard, Michel Beaudouin-Lafon, Yangzhou Du, Caroline Appert, Jean-Daniel Fekete, and Olivier Chapuis. 2006. Shakespeare's complete works as a benchmark for evaluating multiscale document navigation techniques. *Proceedings of the 2006 AVI workshop on BEyond time and errors novel evaluation methods for information visualization* - BELIV '06, 1. http://doi.org/10.1145/1168149.1168165