Of Course I Wouldn't Do That in Real Life: Advancing the Arguments for Increasing Realism in HCI Experiments

Letitia Lew
Truc Nguyen
Stanford University Computer
Science Department
353 Serra Mall, Gates 261
Stanford CA, 94305 USA
{lletitia, nguyen90}@stanford.edu

Solomon Messing
Sean Westwood
Stanford University
Communication Department
450 Serra Mall, Building 210
Stanford, CA 94305
{messing, seanjw}@stanford.edu

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
We offer a nuanced examination of the way that realism can impact internal and external validity in HCI experiments. We show that if an HCI experiment lacks realism across any of four dimensions—appearance, content, task and setting—the lack of realism can confound the study by interacting with the treatment and weakening internal or external validity. We argue furthermore, that realism can be increased while still maintaining control: analogue experiments allow researchers to conduct experiments in more ecologically valid environments and online experiments bridge the gap between the cleanroom and field. While increasing the level of realism in an experiment can introduce noise, technological developments have made it easier to collect rich analytics on behavior and usage.

Keywords
Methodology, realism, experimentation, design.

ACM Classification Keywords
H1.m. [Information Systems]: Models and Principles - Miscellaneous.

Introduction

Experimentation gives researchers the power to reveal causal relationships between variables, test hypotheses and understand new truths. There are two primary types of HCI experiments: 1) *Research Experiments* for which the goal is to explore theories of interaction and produce new knowledge, and 2) *Design Experiments*, for which the goal is to evaluate interfaces and produce new designs [6]. In research experiments, researchers take pains to ensure experimental validity, but do not always use realistic interfaces as the medium for interaction. In user studies for design experiments, designers emphasize rapid prototyping and iteration to refine their product, but tend to conduct the studies in contrived settings, using prototypes that are only partially functional.

Often, both Research and Design experiments lack realism. The degree of realism in an experiment, typically referred to as "ecological validity," describes how closely the appearance, content, methods and setting of the experiment approximate the real-life situation that is being investigated. A widely accepted body of literature has argued that ecological validity is not necessary for experiments in HCI and social science to be valid [3,9], unlike the other four kinds of experimental validity (statistical, internal, construct and external). Below, we will re-examine the impact of realism on experimental validity, and show that a lack of realism can threaten both internal and external validity.

External Validity in HCI Experiments

Lack of realism in the experimental interface or setting can undermine external validity when the level of realism is held constant across all cells of the experiment. The interface and setting are background factors in the experiment, i.e. factors that we do not explicitly manipulate [3]. Sometimes a background factor is not merely "noise", but can interact with the treatment (Fig. 1) to mask or exaggerate (i.e., moderate) its effect [9]. Consider a hypothetical *design* study examining whether users of an iPhone video game enjoy a novel feature, wherein the type of monsters they encounter depends on where they are in their physical world—so that if users are in a park, they will encounter tree- or squirrel-like monsters; in a store, monsters that look like gifts, etc. Suppose the researchers arm players with iPhones to roam through laboratory areas set up to look like a park and a department store, with different monsters appearing in each area for the treatment condition, while in the control condition any monster can appear in any area. We would expect players not to find that the location feature adds much excitement in these contrived settings. If so, the lack of realism will interact with the treatment (Fig. 1) to produce results that disparage the location feature. Since the level of realism is held constant across both conditions of the experiment, we will not be able to replicate our results if we repeat the experiment at a different level of realism, say if players use the iPhone in the city and really get a thrill out of discovering arboreal creatures as they walk by park areas, relative to simply encountering random monsters throughout the city. The experiment conducted in the lab setting is not externally valid.

On the other hand, lack of realism compromises internal validity when the level of realism varies across the treatment and control conditions of the experiment (i.e., mediation). Consider a Research study investigating the effects of displaying advertising on a

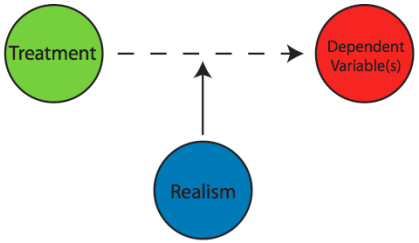

Fig 1. Moderation Diagram

retail website. Say the researchers build an amateur website with uninspiring merchandise displays, and use Flash banners for the treatment condition (with no banners for the control condition). Then we might expect the lack of realism in the website to cause the advertisements to distract subjects much more than they would normally, because the high-quality Flash advertisements stand out against the unprofessional site appearance. As illustrated in Fig. 2, our unusual result of extreme distraction is caused primarily by the realism of the banners in the treatment condition, not by the treatment itself (the presence of the banners). As the experiment lacks internal validity, it is consequently also not externally valid.

We call for greater realism in experimental interfaces to improve the validity of the results. In the next section we will show how lack of realism in an experiment's appearance, content, task and setting can interact with the treatment in both Research experiments as well as Design experiments. By addressing these background factors, we can better ensure that observations of the experiment's dependent variable(s) are not confounded.

Fig 2. Mediation Diagram

Realism Across Dimensions of an Experiment

We will discuss what realism means in various aspects of the design of an experiment, including: (1) the interface appearance and the content of stimulus materials, (2) activity task, and (3) experimental setting. Then we will show how lack of realism in each dimension interacts with the treatment effects and hence threatens validity if held constant. Finally, we will discuss emerging technologies and examples of improving realism in that dimension.

Realism in the Appearance of the Interface

A lack of realism in an interface's appearance or content signals users/participants that they are being tested. This can influence participants' mindsets by reducing their interest or pleasure, or it may cause distraction and curiosity. All of these can influence the user's interaction with the manipulated treatment. Interfaces used in *Research* experiments often lack realism in appearance and content, if researchers only have time to hack together a rudimentary tool for testing. Similar improvements are necessary in early stages of *Design* experiments, where designers create crude mock-ups for prototyping.

Realism in appearance refers to how closely the visual qualities of the experimental interface resemble those of the interface as encountered in real life. These qualities may include its degree of aestheticism, professionalism, graphic representation of mental models, and navigation scheme. All of these inform the user about the interface they are using, and shapes his or her response. For example, unrealistic graphics in driving simulators evoke very different driving behavior than real-life driving. The dull colors and unnatural movements of pedestrians should be expected to decrease drivers' awareness of foot traffic; if so, it is likely that treatment conditions with distractions will show many more accidents in driving simulators than under real road conditions.

It is becoming increasingly feasible to address these shortcomings in realism. Due to the incredible advances in video game graphics, realistic driving simulators have fallen in cost to $20,000 while employing the stunning graphics of game software platforms such as rFactor [14]. Additionally, with the ubiquity of software

development tools and designers who know how to use them, it is fairly easy and common to create platforms and websites that look and function like authentic modern software, using templates and standard layouts. With little expense of effort one can even create spoof websites such as "Amazonaccounts.net" that look and navigate exactly like the original [5]. We also have greater access to previous studies on aesthetics to improve the visual effectiveness of our designs. For example, the Eyetrack study found that users' eyes tend to fixate on the upper left corners of screens first [11], so researchers can adjust their screen layout to draw attention to the elements of interest.

Realism in the Content of an Interface

Even if the technology in a study looks perfectly realistic, users may notice that the system feels contrived if it does not generate the appropriate content. In a user study, participants usually interact with the system to obtain some sort of data: for example, users might be asked to modify a file, or complete a search task. Most HCI experiments maintain control over content by allowing the user to interact only with a set of data which may be fabricated, out-of-date or a biased subset of the full range of possibilities, all of which can feel artificial to the user. Unrealistic content may raise suspicions among participants about the authenticity of the system and may lead to atypical user behavior, such as users attempting to explore the boundaries of the system out of curiosity. In a study to investigate how the race of a news story's protagonist affects news selection, researchers created an 'experimental' news website with each article featuring a picture of either an African-American or Caucasian protagonist, or a non-portrait control picture. However, all the articles were edited and had generic topics such as "eBay quest for cookie jar leads to long-lost sibling" [8]. This lack of realism is held constant, but could interact with the treatment to cause a 'novelty' factor where participants have different motivations in an unrealistic system than in real life. For example, African-American participants may recognize that the articles are fake, and click on more articles featuring African-American protagonists out of curiosity for what the researchers wrote about their ethnic group in the manipulated articles, though they would not feel compelled to do so with authentic news. Thus this study lacks external validity.

Today it is relatively easy to implement search engines to produce actual search results, or RSS feeds to grab current news articles. If the experimental system is slow and real-time performance is a concern, one option may be to fetch all the real content at the start of some suitable timespan and cache the data for that experimental session. This is still better than creating fake content.

Realism in Task

Researchers often conduct usability testing by asking human subjects to perform a series of tasks in an interface. Such tasks ideally include the range of activities that users will actually perform, and allow users enough time to perform these activities [15]. The personal relevance of a task is key to understanding user experience because having users dispassionately "go through the motions" while they are not interested in the task can create a "boredom effect" [12] that could be confused with treatment effects. If task scenarios are contrived, assigned at the wrong level of granularity, or do not engage the user, then these

factors can interact with treatment effects by restricting the user's actions or causing an attrition effect.

Consider a study where a lack of realism in task will compromise external validity. Bergman et al. tested their file "demotion" system by asking users to clean up files in folders, a realistic everyday activity [2]. In the control condition they only had the 'delete' option, and in the treatment condition participants had the additional option to 'demote', or gray out, the file. If the researchers had asked participants to tidy up a generic set of folders, we might expect that participants would delete most files and make little use of the file demotion interface, because the sample files would not matter to them. Thus the lack of personal relevance would interact with the treatment condition and compromise external validity. Fortunately, Bergman et al. conducted the experiment on participants' own PCs to organize their personal files, thus achieving realism in task.

Realism in Setting

A realistic setting is a natural environment in which a user regularly interacts with an interface. Elements of this environment frame the user's physical and mental state, including lighting, the familiarity of the surroundings, the user's physical position with respect to the interface, and the level of noise and bustle from the surroundings. Yet most HCI experiments are still conducted in the laboratory, and the artificiality of such settings can cause behavioral changes in users that confound the treatment effects. The results do not generalize to the real world because the participants' inauthentic responses are only observed in the lab setting. (Recall the iPhone game from Section 1, which did not realize the full enjoyment potential of location-based monsters when the game was tested in the lab.)

The most realistic setting for any experiment is the field, or the actual environment where the interface is or will be used. To document an effect with a field experiment serves as a powerful indicator of the effect's real-world significance, suggesting that the effect is still noticeable in light of the noise and externalities associated with the real-world. For example, in their analysis of the effect of a "mobile awareness system," Oulasvirta, Petit et al. [10] activated a feature that provided users with real-time cues of other people's current context and undertakings, which could be utilized to infer availability of other users. Using an A-B (pre-post) intervention methodology, they provide experimental evidence showing that using the system resulted in a higher rate of call completion (success). Despite of the noise of real world smart phone use, the authors still documented an effect, suggesting strong external validity and generalizability.[1] However, field experiments only produce insights if the experimenter can maintain control and experimental validity throughout.

[1] The authors note that "An A–B intervention methodology from clinical medicine and clinical psychology was utilized." Usually, this methodology is employed when there are ethical issues involved in creating a control condition, which would require the experimenters to withhold treatment. Such pre-post designs do not isolate causality in the same way that random assignment to treatment and control does—and indeed there are external events in this study that might not be orthogonal to the intervention (treatment), such as summer break. We recommend simple random assignment to treatment and control in the context of HCI, where the ethical issues discussed above are seldom present.

One way to approach realism while maintaining control is to decorate the lab in likeness of the field. These "analogue experiments" are more realistic than lab experiments but more reliable than field experiments. An effective analogue experiment has the potential to expose all the usability issues that arise in a field experiment for *Design* purposes. Kjeldskov et al. tested a mobile medical device in both an actual hospital and a lab decorated to look like a hospital, and found that their analogue experiment uncovered more important usability issues than the field experiment did [7]. Current research does not show how the two compare for *Research* experiments. The advantage of analogue experiments is that we need not surrender control to gain insights on how users naturally interact with interfaces.

Online experiments are beginning to bridge the gap, as they can be executed with the ease of analogue experiments, but approach the realism of field experiments. As use of the Internet grows more widespread and many common tasks move online, the setting in which users choose to encounter the treatment is also more realistic. Shifting towards web experiments can be an affordable and valid way to improve realism in setting. The Internet offers portability and ubiquity and can be used to reach a more heterogeneous sample of participants, overcoming the problem of biased sub-populations [1].

Issues with Increasing Realism

While increasing realism reduces the threats that *artificiality* poses to external validity, it can introduce problems with other aspects of experimental validity. Using realistic content and environments means that we relinquish some degree of control to let users interact naturally with elements of the interface. When considering the realism of a study, considerations of both positive and negative consequences of realism are critical in achieving an ideal level of balance between realism and experimental control. Researchers must be careful not to shoot themselves in the foot, as Lynch points out that we "cannot enhance the external validity of our study by making it more realistic if the methods used to increase realism threaten the internal validity of the study" [9].

To achieve credible experimental results, we must consider the threats of increasing realism to the internal, external and construct validity of our experiment. We will then explore how to mitigate these risks in the initial design, and how to incorporate tests of validity in the experiment for subsequent revisions.

Problems with Respect to Internal and External Validity

Increasing realism can weaken internal validity because it introduces more noise into the data collected, making it trickier to demonstrate causal relationships between variables in an experiment. When the experiment utilizes more realistic content and settings, there is greater risk of random irrelevancies [3] that muddy the data. In the realistic scenarios we described in Section 2, participants may encounter unexpected disruptions or topics close to their hearts, which may alter their behavior and obfuscate a causal link between variables. Another problem is that realistic settings diminish the reliability of treatment implementation. Participants may not hear instructions in the field, or data logging devices such as videocameras may miss events.

Increased realism may also pose problems to the external validity of an experiment by causing other

factors (besides artificiality) to interact with the treatment. One such interaction is that of history and the treatment [3, 10] where the results obtained on a particular day may not generalize to other days. Consider a web experiment examining the effects of no-minimum-purchase free shipping on consumers' shopping behavior. The experiment involves consumers shopping for personal items in real online stores, exhibiting realism in content, task and setting. But if this experiment were conducted during the holiday season, where consumers typically buy gifts for many different households, then we would expect that they would buy far more with free shipping than they would at any other time of year. This data is not externally valid across all time periods.

The interaction of testing and the treatment may also reduce external validity. If the methods used to capture field data are obtrusive, for example with visible equipment such as videocameras, then the user is constantly aware of the fact that they are being tested, which can cause subjects to behave differently than they might otherwise (i.e., the "Hawthorne effect"). A related phenomenon is 'evaluation apprehension', where the user feels pressure to do their best in an expensive field test.

One solution to these problems with internal and external validity is to measure everything, and to do so in the background. If researchers collect data on every aspect of the user's interaction with the interface, they will be better equipped to trace how "the uncontrolled factors could have caused local departures from the modal effect" [4] in later analysis. For example, a study on an expertise search system, SmallBlue, used realistic content that had the potential to compromise its internal validity [13]. Employees of a big corporation used the SmallBlue system to search for coworkers who were experts in AJAX. This could be a problem if particular experts were well known across all participants, causing familiarity with an expert to override the other indicators of expertise being studied. The researchers kept track of whether users directly knew one of the colleagues returned by the search, and factored this into their 'social closeness' parameter. They were able to statistically control for this factor in follow-up analysis.

Increasingly, background data-logging can be done cheaply, with little prior expertise and experience. Computers and web-interface platforms can now record a variety of measures unobtrusively, including file revision histories, screen sharing, integrated videorecording, and mouse movement tracking. This affords researchers finer measures describing the content the user is viewing, what actions they take, and can even provide indicators of attrition and disruption. This provides researchers with greater opportunities to detect any interaction effects and appropriately handle the data points whose internal or external validity has been compromised. Much of the software available for this purpose is open-source and/or cheaply available for use in academic research settings.

Implications for Measurement, Reliability and Construct Validity

Compared to the relatively common practice of issuing user questionnaires or interviews, having users perform real tasks and measuring how they actually behave translates into far less ambiguity and thus finer measurement. Not only does this translate to greater reliability, but the greater correspondence between the

theoretical concept and actual behavioral measures means that construct validity is improved as well. Increasing experimental realism generally *strengthens* construct validity.

The background logging methods mentioned above measure user interaction in a more realistic way, as it measures what actions users actually perform. Contrast this with questionnaires, which only provide "self-reported exposure" to the causal variable in question [1] but are far more widely used as they are easier to implement. For example, when estimating initial demand for a new product, instead of asking participants whether they would buy it, it would be more revealing to set up an online store and count how many of them actually click to purchase. This creates a more realistic shopping scenario, which also maps more closely to our construct of demand.

We recommend a multi-method approach, combining behavioral logging with questionnaires, to supplement the behavioral data with information on users' perceptions. This can alleviate the problem of mono-method bias. With multiple measures of the same construct available to the researcher, a variety of confounds are more easily detectable: construct validity issues, question-wording problems, data-logging errors, and other design flaws. The benefits of increasing construct validity in an experiment often outweigh the risks of increasing realism to internal and external validity.

Conclusion

We have shown how the lack of realism can confound an HCI study by interacting with the treatment and weakening both internal and/or external validity. A lack of realism can *mediate* the treatment, masking or exaggerating its effect; or it can *moderate* the treatment, so that the treatment effectively causes a perceived increase in realism, which is then causes the observed effect. We have demonstrated how a lack of realism across any of four dimensions—appearance, content, task and setting—can interact with the treatment and confound the experiment.

Experimentation is an important scientific tool through which to advance knowledge in the field of HCI. We have identified common problems in both *Research Experiments* that aim to test HCI theories, and *Design Experiments* that aim to develop new interfaces. Experiments of the former type tend to lack realism in the appearance and content of the interface, while the latter type tend to put more effort into making the interfaces realistic, but experimental tasks and settings are less carefully planned.

The solution we propose is to make interfaces in both types of experiments more realistic. Current research on the effects of realism in HCI experiments only considers realism in a single aspect of the experiment, for example only the setting [10]. We have shown that increasing realism in all dimensions can strengthen experimental validity in the following ways: enhancing realism in *appearance* and *content* makes the experimental interface in a Research experiment more comparable to real-life technology. On the other hand, user studies for Design experiments often benefit from improving realism in *task* and *setting*, which allows researchers to better contextualize relationships between humans and interfaces and leads to more externally valid results. We have listed exemplary

experiments and technology to aid researchers in improving the realism of their own experiments.

We recognize the risks of increasing realism and show how to counter them by using more realistic data collection methods. Increasing realism tends to introduce more unpredictable disruptions into the data, but measuring all the user's interactions using background logging technology not only enables researchers to monitor for internal and external validity, it also enhances construct validity.

We conclude that enhancing realism in experimental design can lead to better research outcomes, especially when researchers pay close attention to realism and validity simultaneously.

While of course not all experiments can justify the time and expense needed to increase realism, we recommend that researchers consider increasing realism for studies where validity is crucial: in Research experiments for academic publishing, and in Design experiments for late-stage products. In these studies the results form the foundation for future academic progress in the former, and for costly business decisions in the latter. Modern interface design and data logging technology also enable us to improve realism in existing experiments and to possibly add insightful results to previous studies that were conducted with minimal realism.

Acknowledgements

We would like to thank Professor Clifford Nass for giving us the opportunity to participate in research that inspired this note and the reviewers for their helpful comments.

References (should be in alpha order)

[1] S. Ansolabehere, S. Iyengar, A. Simon, and N. Valentino, "Does attack advertising demobilize the electorate?," American Political Science Review, 1994, pp. 829–838.

[2] O. Bergman, S. Tucker, R. Beyth-Marom, E. Cutrell, and S. Whittaker, "It's not that important: demoting personal information of low subjective importance using GrayArea," Proceedings of the 27th international conference on Human factors in computing systems, 2009, pp. 269–278.

[3] Cook, D.T. Campbell, and A. Day, Quasi-experimentation: Design & analysis issues for field settings, Houghton Mifflin Boston, 1979.

[4] L. Cronbach, "Beyond the two disciplines of scientific psychology," American Psychologist, vol. 30, pp. 116–127.

[5] S. Egelman, L.F. Cranor, and J. Hong, "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, 2008, pp. 1065–1074.

[6] D. Fallman, "Design-oriented human-computer interaction," Proceedings of the SIGCHI conference on Human factors in computing systems, 2003, p. 232.

[7] J. Kjeldskov and C. Graham, "A review of mobile HCI research methods," Human-Computer Interaction with Mobile Devices and Services, 2003, pp. 317–335.

[8] S. Knobloch-Westerwick, O. Appiah, and S. Alter, "News Selection Patterns as a Function of Race: The Discerning Minority and the Indiscriminating Majority," Media Psychology, vol. 11, 2008, pp. 400–417.

[9] J.G. Lynch Jr, "On the external validity of experiments in consumer research," Journal of Consumer Research, vol. 9, 1982, pp. 225–239.

[10] A. Oulasvirta, "Field experiments in HCI: promises and challenges," Future Interaction Design II, 2008, pp. 1–30.

[11] S. Outing and L. Ruel, "The best of eyetrack III: What we saw when we looked through their eyes," Published on Poynter Institute (not dated). Retrieved, vol. 20, p. 06.

[12] W. R. Shadish, T. D. Cook, and D. T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, 2nd ed. Wadsworth Publishing, 2001.

[13] N. S. Shami, K. Ehrlich, G. Gay, and J. T. Hancock, "Making sense of strangers' expertise from signals in digital artifacts," in Proceedings of the 27th international conference on Human factors in computing systems, pp. 69-78, 2009.

[14] G. Weinberg and B. Harsham, "Developing a low-cost driving simulator for the evaluation of in-vehicle technologies," Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2009, pp. 51–54.

[15] C. Wharton, J. Bradford, R. Jeffries, and M. Franzke, "Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations," in Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 381-388, 1992.