

The Effect of Parallel Prototyping on Design Performance, Learning, and Self-Efficacy

Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Scott R. Klemmer

Stanford University HCI Group
Department of Computer Science
Stanford, CA 94305

[spdown, alana11, jkass, schwarz1, srk]@stanford.edu

ABSTRACT

Designs often improve with iteration. Does creating and receiving feedback on multiple prototypes in parallel— as opposed to serially— affect outcome, learning, and self-efficacy? Two experiments manipulated whether participants designed prototypes and received feedback in parallel or serial. In the first, participants designed Web advertisements and received descriptive critique on each prototype. As measured by click-through rate and expert ratings, ads created in the *Parallel* condition significantly outperformed those in the *Serial* condition. Independent raters found Parallel prototypes to be significantly more divergent. Parallel participants reported a larger increase in task-specific self-confidence. In post-task interviews, several Serial participants reported negative reactions to critique; no Parallel participants reported this. The second study manipulated Parallel versus Serial for a simple mechanical design task. This study found no significant performance difference between conditions. We discuss differences between the tasks and the implications for understanding when and how parallel prototyping is beneficial.

Author Keywords

Prototyping, iteration, feedback, juxtaposition, design

ACM Classification Keywords

H.1.m. [Information Systems]: Models and Principles

General Terms

Experimentation, Design

INTRODUCTION

Iterative prototyping is central to learning and motivation in design [19,23,32,48,49]. Iteration's virtue— incremental, situated feedback— can also blind designers to other alternatives, steering them to local, rather than global, optima [15,33]. To combat this, creating multiple alternatives in *parallel* may encourage designers to more effectively discover unseen constraints and pliable variables [19], enumerate more diverse solutions [15], and obtain more authentic and diverse feedback from potential users [52].



Figure 1 This research manipulates how participants receive feedback during a design process: Serial (top) versus Parallel (bottom).

Even if parallel prototyping has these effects, it necessarily takes time away from refinement. This paper investigates the relative merits of parallel and serial prototyping under time constraints.

Two experiments manipulated whether participants designed prototypes and received feedback in parallel or serial. In both, participants created five prototypes and a final design within the same overall time period. In the *Serial* condition, participants received feedback on one prototype at a time. Participants in the *Parallel* condition created three prototypes, received feedback on all three, then made two more prototypes, and received feedback again (see Figure 1).

In the first study, 33 participants designed Web banner advertisements for a magazine. Participants received a descriptive critique on each prototype. Final advertisement performance was measured through online data analytics from a MySpace.com ad campaign and expert ratings. A measure of divergence was obtained through independent online raters who judged pair-wise similarity between each of the participants' six ad designs.

Stanford Tech Report
September 2009

In the second study, 35 individuals created a mechanical “egg-drop” vessel from everyday materials, designed to protect a raw egg from a fall. Participants tested each prototype by dropping it. The height at which the egg breaks is the dependent variable.

Both studies employ tasks with the following properties: success is objectively measurable, participants need minimal technical or artistic expertise, there are many possible valid solutions, and the activity can be completed within a single session. Both studies administered a pre- and post-test self-efficacy [27,31] and concluded with an open-ended interview.

In the ad study, Parallel participants outperformed Serial participants by all performance measures: click-through rates, time on client site, and client and ad professional ratings. Independent raters found the set of a participant’s prototypes to be more dissimilar in the Parallel condition. In post-task interviews, several serial participants reported negative reactions to critique of their prototypes; no Parallel participants reported this. Parallel participants also saw a significant gain in self-efficacy scores—a measure of confidence on the specific task; Serial participants did not.

The egg-drop study yielded no performance difference between the Parallel and Serial condition. The study manipulation did not affect egg-drop participants as it did in ad design. Test drops are ephemeral making explicit comparison difficult regardless of condition. Likewise, the first study’s finding of negative reaction to critique did not pertain to objective egg drop trials. The differences between the ad design and egg-drop design tasks shed light on when and why parallel prototyping affects design outcome.

DESIGN STRATEGIES

In design, problems and solutions co-evolve, constraints are often negotiable, sub-problems are interconnected, and solutions are “not right or wrong, only better or worse” [22,29,37,48]. Descriptions of design generally feature exploration, refinement, and iteration [8,9,19,23]. Some accounts foreground the role of formal models [8,26], others a trial-and-error approach [15,35,49,53].

Successfully navigating the “wickedness” of design problems requires balancing concerns [46]. One danger is to refine too early and fail to identify a valuable direction [12,18]. Designers may “fixate” on initial ideas [24,34], make poor choices to justify prior investments in money or time [7], or make only “iterative improvement of the same design” [52]. Conversely, too much exploration and there isn’t enough time to execute [10,12]. Without refinement, ideas may not reach their full potential [12].

Laseau posits an idealized model for exploring and refining, where designers iteratively diverge and converge on ideas, eventually narrowing to a best-fit concept [39]. This paper experimentally investigates this theory by contrasting a Parallel “explore-then-refine” prototyping strategy with Serial refinement.

Does Parallel Feedback Promote Learning in Design?

Our intuition says parallel prototyping encourages comparison among multiple divergent alternatives. Throughout life, people learn by interacting and correlating actions with perceivable differences in the world [30,44]. Life experiences provide a corpus of examples from which to draw comparisons in new learning situations [36]. Examples aid in problem solving [5,50] and provide greatest benefit if people explicitly extract principles, facilitating more effective application [28,51]. Comparison helps people focus on key relations, aiding the acquisition of underlying principles and sharpening categorical boundaries [14,17]. Comparison’s value provides a rationale for parallel prototyping: designers more effectively discover key variables and their interrelations, which produces actionable steps forward.

Hypothesis 1: *Parallel prototyping encourages more dissimilar designs.*

Hypothesis 2: *Parallel prototyping produces higher quality designs.*

This paper investigates these hypotheses by measuring task performance, comparing the diversity/similarity of prototypes, and by coding interview data.

Does Parallel Feedback Affect Motivation and Self-Efficacy?

Motivation helps designers face the challenges, setbacks, and uncertainty inherent in prototyping [35,49]. Without motivation, designers may settle for mediocrity (“it’s good enough”) [11] or perceive feedback that only confirms their ideas (i.e., confirmation bias [42]). Motivation and self-efficacy improve one’s ability to learn, perform towards a goal, exert agency, persist, and find enjoyment in challenges [20,25,41,43]. Monetary rewards are one motivator, but they have been shown less effective on creative tasks [6,45]. Other research says people engage and perform in activities to satisfy a desire such as curiosity, autonomy, or mastery [40,45]. Csikszentmihályi claims optimal motivation occurs when an activity strikes the right balance of challenge and abilities [20]. We propose that parallel prototyping—by providing opportunity for comparing feedback and combining elements—instills intrinsic motivation for action and progress.

People who believe they can perform well are more likely to view difficult tasks as something to be mastered rather than something to be avoided [13]. People with high self-efficacy respond less negatively to failure, focus on strengths, and repress weaknesses [21]. In design critique sessions, or “crits,” self-efficacy mediates how feedback is interpreted. People can conflate honest criticism as a personal judgment rather than an assessment of the concept itself [38]. The studio model of art and design education tries to minimize these effects by framing critique in terms of “the work” rather than the person [47]. Furthermore, as Tohidi et al. showed, the presence of multiple alternative concepts enables reviewers to be more critical with their comments [52].

Parallel feedback on multiple prototypes provides designers an opportunity to engage in comparison and potentially avoid discouragement caused by subjective critiques.

Hypothesis 3: *Parallel prototyping yields increased self-efficacy.*

Hypothesis 4: *Serial prototyping yields increased frustration.*

This research investigates these hypotheses with a task-specific test of self-efficacy (administered before and after the task) and by coding interview data.

RESEARCH METHOD

The studies described in this paper manipulate the structure of the prototyping process. In one study, participants designed Web advertisements; in the other, participants created an “egg-drop” device [1]. The studies were designed and administered concurrently. The tasks were selected to satisfy the following criteria:

- Quality can be measured objectively and subjectively;
- Subjects need minimal artistic or engineering ability;
- Individuals can complete tasks within a lab session;
- Solutions demonstrate creative diversity and perform at a range of levels;
- The study procedure could generate fair and authentic feedback during iteration.

Both studies hold constant the number of prototypes created, the amount of feedback provided, and the overall time allotted. In the *Parallel* condition, participants create 3 prototypes and get feedback, then 2 more, then a final version. In the *Serial* condition, participants create 5 prototypes in series, receiving feedback after each prototype, then a final.

Study 1: Web Advertisement Design Task Instrument

Subjects design a 160×600 pixel banner ad to be hosted on the social networking site MySpace.com. Ads were created using MySpace’s Web-based *AdBuilder* tool (see Figure 2). This simple graphic design tool was easy to learn and no participants had used it before (which prevented giving an advantage to participants with particular tool skills.)

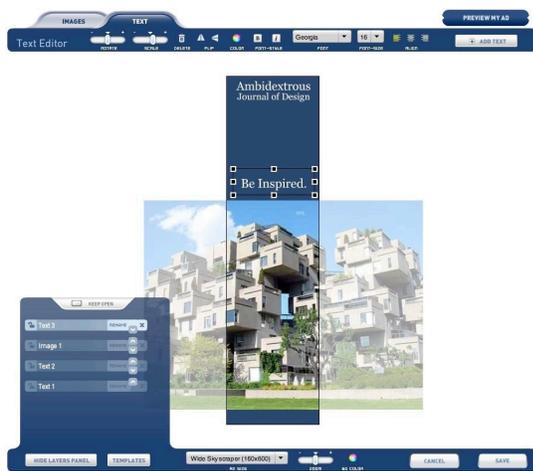


Figure 2 The ad design study used MySpace’s *AdBuilder*, a browser-based graphic design tool

Prior to the experiment, a team of three advertising and graphic design professionals developed a list of 50 statements that could serve as critique for any banner ad. The list included three categories of statements—overall theme, composition & layout, and surface elements. Each category contained 12 to 20 statements (see Appendix A). During the experiment, the experimenters chose three statements—one from each category—to attach to each ad prototype. Critique statements were intended to provide high-level direction, without using explicitly positive or negative language. Importantly, the feedback was the same in both the conditions; the *Parallel* condition provided no explicit comparisons between multiple ads, such as, “The color in this ad is better than that one.” *Parallel* feedback critiqued each ad independently, matching the stimuli in the *Serial* condition.

After the experiment all final 33 ad designs were uploaded to MySpace for a 15-day campaign targeted to users interested in design-related activities. Design performance was measured through four dependent variables:

- MySpace reports: daily number of clicks and impressions (number of appearances on MySpace);
- Google Analytics [2] on the client Web site: number of clicks, time spent on site, and number of pages visited for each ad on each day;
- Client ratings (four editors of *Ambidextrous* [3]);
- Expert ratings (three Bay Area ad professionals).

The four editors of the magazine and three advertising professionals rated the participants’ ad designs from 0 to 10 along five dimensions: adherence to the client’s theme, creativity/originality, visual appeal, tastefulness, and adherence to graphic design principles. Raters were blind to condition and rated ads individually, with no knowledge of other raters’ scores.

Workers on Amazon’s Mechanical Turk [4] rated pair-wise similarity between each of the participants’ six ads (see Figure 3). Workers were randomly presented a pair of ads from a participant and asked to judge their similarity on a scale from 0 to 7 (not similar to very similar). This measure generated 14,850 judgments (30 worker assessments on each of the 15 pair-wise comparisons for 33 participants).



Figure 3 Example pair-wise ad similarity rating

	AMBIDEXTROUS <small>Stanford University's Journal of Design</small>	Subject ID # <u> 137 </u> Prototype # <u> 2 </u>
	Ambidextrous wants an ad that reaches out to design practitioners, students, and researchers.	Try to create a visual flow for the viewer—what should the viewer see first?
	Use color to create emphasis, to separate different elements, or to categorize content.	<small>Quarterly Print Magazine 3rd Year of Publication</small>

Figure 4 Example feedback form

Procedure

The ad task proceeded with the following parts: consent form, demographics, instructions, practice task, design brief, prototyping period (10 minutes per prototype), final design period (15 minutes), questionnaires, and interview. The design brief detailed the ad campaign's client, *Ambidextrous* design magazine [3] and outlined three goals: increase traffic to the *Ambidextrous* Web site, impress the editors, and create ads with effective graphic design.

Participants were instructed they would receive feedback critique from an ad expert on each prototype. As experimenters prepared critique forms in a separate room, participants were allowed to navigate the client Web site, search for images, or sketch on paper. After five minutes, participants received an envelope containing the printed ad prototype with feedback statements (see Figure 4). For 150 minutes of participation, subjects received \$30 cash. Experiment proctors only entered the same room as the participant during the introduction and instructions, and to deliver feedback envelopes.

Study 2: Mechanical Egg Drop Exercise

Instrument

In the *egg drop exercise*, participants designed a vessel from everyday materials to protect a raw egg from a fall. The following design materials were available: 8 pipe cleaners, 8 rubber bands, 8 popsicle sticks, one 4"×8" piece of poster board, one sheet of tissue paper, one 4"×6" piece of flat foam, and one foot of scotch tape (see Figure 5). Pilot studies showed the choice of materials to be diverse enough to elicit many approaches yet challenging enough to produce a wide range of performances (More materials result in top performers too high to be easily tested). Performance is determined by dropping a single egg from a one-foot marker,



Figure 5 Materials for Study 2: pipe cleaners, wooden sticks, rubber bands, tissue paper, poster board, and flat foam.

then two, then three, and so on until the egg cracks. During the testing periods, participants could test their prototypes in an area adjacent to the worktable. For the final official drop, a member of the research team repeatedly dropped the vessel to determine the survival height (recorded in feet).

Procedure

Participants first filled out a consent form and demographics questionnaire. The experimenter verbally described the exercise and the specific rules for the assigned condition. All participants were told they would have 60 minutes to create and test five prototypes (12 minutes per prototype). They were given a set of construction materials, and were told they could get replacement materials if necessary. After the prototyping period, the researcher cleared the workspace and provided a fresh set of the original materials (this time without replacements). Participants were given 12 minutes to build the final design, followed by a 10-minute questionnaire, a 10-minute interview, and the egg drop test. The short open-ended interview asked participants to describe their concept, process, and theories about how the egg might break. For 120 minutes of participation, subjects received \$25 cash.

Method Common to Both Studies

Questionnaires

Both studies administered pre- and post- design task questionnaires and interviews. Participants were asked to list out variables/factors believed to be important in the respective tasks (e.g., "list all factors that make a good Web ad"), both before and after the design task. After the design task, participants critiqued another ad or egg drop design. They also filled out the "Creativity Achievement Questionnaire" developed by Carson et al. to assess creative achievement across ten domains [16].

A task-specific measure of self-efficacy assessed an individual's belief in their ability to perform the task (adopted from self-efficacy exams in education [27,31]). The exam asks participants to rate their ability to: create advertisements [or egg drop vessels], understand design problems, detect problems in a design idea, and incorporate feedback into a design idea. The 7-point Likert scale for each question provided an overall range between 0 and 24. The same questions were administered before and after the design task, creating a difference measure. The change in self-efficacy score is important as it provides an indication of how condition (Parallel/Serial) and other factors influenced an individuals' belief in their design abilities.

Participants

Participants were recruited locally with fliers and randomly assigned to one of two conditions. Both studies balanced for gender and task relevant experience. In the ad design study (N=33), there were 19 females and 14 males. Fourteen participants reported some prior experience in ad or graphic design. In the egg drop study (N=35), there were 19 females and 16 males. 17 had previously taken part in the egg drop exercise.

PERFORMANCE RESULTS

The primary independent variable (Parallel versus Serial) had a significant effect on performance in the ad design task, but not in the egg drop task.

Advertisements

Online Performance Data

Performance data on each ad was extracted from MySpace and Google Analytics on the *Ambidextrous* Web site (see Table 1). MySpace reports that over the 15-day campaign, the 33 participant ads received 501 total clicks on 1,180,320 total impressions (i.e., number of ad appearances), giving an overall average click-through rate of 0.0424% or 424 clicks per million impressions (CPM). The top two click-through rates were both Parallel ads with 735 and 578 CPM, respectively. The bottom four ads were all from the Serial condition, with two ads receiving no clicks at all.

It is important to note that, like many advertising hosts, MySpace varies the number of impressions based on prior performance of the ad¹. MySpace's algorithms likely improve user experience and revenue, but they complicate the task of comparing ad campaigns. Parallel ads were shown more than Serial ads, but click count alone cannot measure performance since some ads are unfairly advantaged with click opportunities (impressions). This differential treatment can be explained by the performance of these ads early on, when the two conditions received an approximately equal number of impressions. After day five, Parallel ads had 79,800 impressions with 44 clicks and Serial ads had 79,658 impressions with 26 clicks; at this early stage, Parallel ads had a significantly higher click-through rate ($\chi^2 = 4.59, p < .05$).

A general linear model analyzed variances in final click-through rates for each ad. Parallel outperformed Serial ($F(1,30) = 4.227, p < .05$). Additionally, at marginal significance, highly creative individuals outperformed low creativity ($F(1,30) = 3.812, p = .06$). Incidentally, since the creativity questionnaire was administered at the end, 11 of the top 16 creative folks happened to be placed in the Serial condition. This suggests the relative performance of Parallel ads could have been even stronger if the study had balanced for creativity.

According to Google Analytics on the client Web site, the site received 422 total visitors during the 15-day campaign, 79 less than the number of clicks reported by MySpace. One possible explanation for the difference between Google Analytic visitors and MySpace clicks could be that users clicked the ad and then hit "back" before the browser

¹ MySpace does not publish their algorithm for determining the frequency of impressions, but a repeated measures general linear model with the Day 5 CTR as a factor and impressions on each subsequent day as dependent measure shows the CTR for days 1-5 to be a significant predictor of the number of impressions for the of the final 10 days of the campaign ($F(1,29) = 23.2$ and $p < 0.01$). MySpace receives payment on each click; intuitively, it is in their interest to show high-CTR ads more often.

	Parallel	Serial
MySpace Data		
Total impressions	665,133 (SE=10992)	515,187 (SE=8822)
Total clicks	296 (SE=6.7)	205 (SE=4.7)
Total clicks per impressions	0.0445%	0.0398%
Clicks per million	445.0	397.9
Google Analytics		
Total visitors	264 (SE=5.0)	158 (SE=3.7)
Total time (sec) on client site	7510 (SE=236.5)	3283 (SE=87.7)
Average time (sec) per visitor	28.4	20.8
Pages visited on site	394 (SE=7.9)	198 (SE=5.1)
Pages visited per visitor	1.49	1.25

Table 1 Summary of campaign data from MySpace and Google Analytics (standard error provided when available).

loaded the client site. The 264 visitors for Parallel ads and 158 visitors for Serial ads are statistically different when compared against impressions ($\chi^2 = 6.61, p < .02$). Were Parallel ads more effective at appealing to the intended audience? Normalizing for visitors, the average time-on-site for Parallel ads (28.4 seconds) was significantly greater than Serial ads (20.8 seconds) ($\chi^2 = 9.06, p < .01$). The result suggests Parallel ads were more likely to reach people genuinely interested in the product offered by the clients.

Effect of Prior Experience on Ad Performance

Participants with prior experience in ad or graphic design significantly outperformed novices. Ads by participants with prior experience received 350 clicks on 752,424 impressions, compared to a ratio of 151 clicks on 427,896 impressions by novices ($\chi^2 = 8.09, p < 0.01$). There was no interaction effect between condition and participants with experience. Again looking at the skewed number of impressions: during the first six days when ads by novices and ads participants with prior experience received approximately the same number of impressions—100,975 and 106,865 respectively—the "experienced" ads received significantly more clicks, 56 compared to 35 ($\chi^2 = 6.11, p < 0.02$).

Visitors also spent far more time on the client's site after clicking ads by experienced participants (28.9 sec/visitor) compared to those created by novices (17.8 sec/visitor) ($\chi^2 = 19.23, p < 0.001$).

Expert Ratings

The average expert rating across all ads was 23.0 out of 50 (35.6 high and 15.0 low). The three top-rated ads were from the Parallel condition. A general linear model with repeated measures analyzed the effects of condition, ad, and rater type (client or professional) on the overall ratings. There was a significant difference ($F(1,5) = 7.948, p = 0.037$) between the average expert rating of Parallel ads ($\mu = 24.4, SE = .92$) and Serial ads ($\mu = 21.7, SE = .83$). There were consistent differences among ads across the seven raters ($F(1,5) = 12.606, p = 0.016$) and no significant difference be-

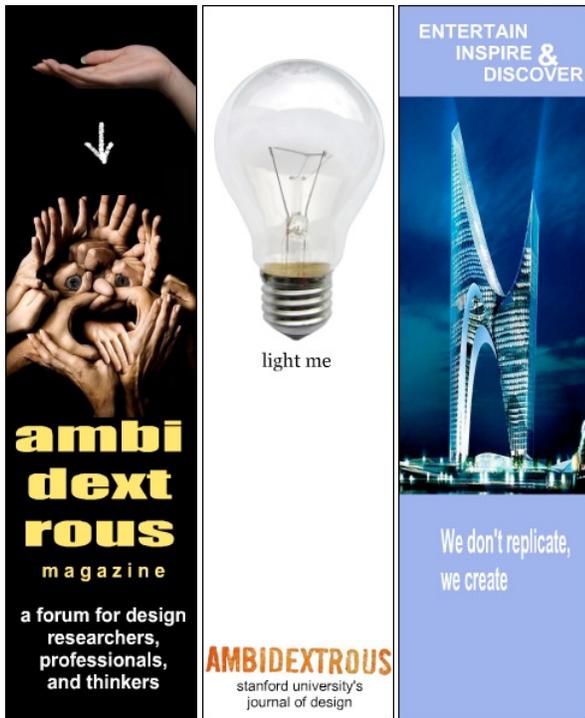


Figure 6 Example ads: (Left) Parallel ad, 1st in click-through rate, 6th in expert rating; (Middle), Parallel ad, 1st in expert rating, 4th in CTR; (Right) Serial ad, 4th in CTR, 32nd in expert rating.

tween ratings of clients and ad professionals ($F(1,5)=0.389$, $p=0.560$), providing a check of inter-rater reliability.

Participants with some prior graphic or ad design experience received significantly higher ratings (25.9) than those with no prior experience (20.9) ($F(1,31)=8.001$, $p=0.008$). The expert ratings revealed no interaction effect between condition and participants with prior experience.

Qualitative Analysis

Questionnaires and open-ended interviews examined how participants explored design possibilities, dealt with feedback, and learned about the principles of ad design.

The ads that performed well online generally also received high ratings by the clients and ad professionals. The ad with the best overall click-through rate received the 6th highest rating by the clients and ad professionals (see Figure 6, left). Likewise, the highest rated ad achieved the 4th highest click-through performance (see Figure 6, middle). The most successful ads (high performance metrics and high ratings) tended to be simple, visually balanced, professional, creative, matched the theme of the magazine and contained some sort of intriguing hook, such as the “face made of hands” in the highest click-through performer.

There were anomalies, such as the top two ads in the Serial condition. These two ads were ranked 25th and 32nd (out of 33) by the expert raters, but for whatever reason received the 3rd and 4th best click-through rates. The latter of those designs does not even mention the client (see Figure 6, right).

Qualitatively, the ads created in Parallel tended to be more divergent than Serial ads (for an example, see Figure 1). Mechanical Turk raters provided a quantitative measure of diversity/similarity. Raters performed pair-wise similarity comparisons on a scale of 0 to 7 within each participant’s set of six ads. Serial ads were deemed significantly more similar than Parallel ads, 3.18 and 2.78 respectively ($F=181.853$, $p<0.001$). The interview data provides additional insight, as one Serial participant explained, “I think the feedback helped. I kept repeating the same mistakes, but maybe less and less each time... the feedback reiterated that.” Another Serial participant said:

I would try to find a good idea, and then use that idea and keep improving it and getting feedback. So I pretty much stuck with the same idea.

This notion of feeling “stuck” or using the feedback to decide where to go next did not surface in the Parallel condition. As one Parallel participants reported: “I didn’t really try to copy off of the ads that I did before...I just made new ideas.” Both the divergence measure and the qualitative interviews suggest the parallel structure reduces the occurrence of functional fixation [34].

In the open-ended interviews, 13 of 16 Parallel participants said the feedback was helpful or intuitive compared to 6 of 17 in Serial ($\chi^2= 3.02$, $p<0.1$). More notably, 8 of 17 of the Serial participants reported the feedback as negative, compared to no such reports in the Parallel condition; this is a significant difference ($\chi^2=7.53$, $p<0.01$). As an example, a participant in the Serial condition said:

I received really negative comments saying [the clients] are looking for a creative and clever ad, which in other words is saying that this is stupid or ridiculous.

More frustration emerged more in the Serial condition, likely because they had no other alternatives on the table. Where a comment about cleverness may offend Serial participants, it provides guidance to Parallel participants; it tells them “one of my other two ideas is headed the right direction.” The same statement is interpreted differently depending on the context.

Did the experimental manipulation affect how participant view the design process? 11 of 16 Parallel participants said on future design projects they would create more than one prototype and obtain copious feedback; only 5 of 17 Serial participants made similar claims ($\chi^2= 2.63$, $p>0.05$). As one Parallel participant said:

Not spending too much time on any single prototype is useful because then you don’t go into details too much.

Another Parallel participant stated that making multiple prototypes was a “great strategy,” but would not implement it because “it’s too much work,” suggesting perhaps motivational factors may prevent some people from adopting an effective process.



Figure 7 In the egg drop-task, craft had a large impact on performance, often dominating the effect of design choices. For example, the best vessel (left, 17ft, Parallel) shared many of the same properties as the worst design (right, 0ft, Serial).

Egg Drop Vessels

Performance Metrics

A two-way analysis of variance was performed with condition (Serial/Parallel) and prior egg drop experience (prior/no-prior) as factors, and height as dependent variable. Condition was not a significant factor ($F(1,31)=0.11$, $p=0.917$). Participants in the Parallel condition reached an average height of 5.4 feet, with those in Serial reaching 5.5 feet. The average height of vessels by participants with prior experience (6.0 feet) versus novices (4.9 feet) is not significant ($F(1,31)=0.703$, $p=0.408$). There was no interaction effect between condition and prior experience.

Qualitative Analysis

The best egg drop vessel was created in the Parallel condition and protected an egg from 17 feet; the worst vessel was created in Serial and failed to protect the egg from a foot. The most successful egg drop vessels were constructed well and accounted for several key variables: slowing the fall, distancing the egg from the first point of impact, handling the balance after impact, and containing the egg. There were exceptions, like the top performer, which made no attempt to slow the fall, but had a clever internal cushion design. The worst design failed because the egg was not secured properly, so it rolled out upon impact. Craft played a big role in performance, often dwarfing the effects of design decisions; the best and worst designs, for example, shared many of the same properties (see Figure 7).

The open-ended interviews revealed a few differences between the conditions. 9 of 17 Serial participants and 3 of 18 Parallel participants said they tended to elaborate on a single idea ($\chi^2=3.36$, $p<0.1$). Similarly, 5 of 17 Serial participants (compared to none in Parallel) said they could not think of alternative ways to design an egg drop vessel ($\chi^2=5.29$, $p<0.05$). In the Parallel condition, participants reported they would intentionally test very divergent ideas, especially within their first three prototypes. As one Parallel participant stated, “I was just trying to make three that were very different, and test different components.” Similarly, another Parallel participant said:

I wanted to make them as different as possible. Maybe one design is better than the other. I don't know how the physics works, so I can't calculate which design will work.

While Parallel participants described a qualitatively more divergent process, they generally reiterated the exploratory prototyping process imposed. Notably, the Parallel feed-

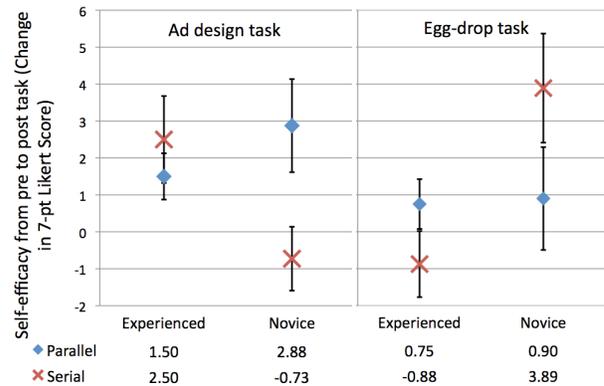


Figure 8 Measure of net gain or loss in self-efficacy scores from before and after the design tasks. Illustrates the interaction effects between condition and the participants' prior experience.

back was not simultaneously comparable since participants had to drop each egg vessel one at a time. These issues are discussed further later in the paper.

SELF-EFFICACY RESULTS

Both studies measured how Parallel versus Serial prototyping process affects self-efficacy (participants' belief in their ability to perform the design task). The self-efficacy questions ask a participant to report their ability to: create advertisements [or egg drop vessels], understand design problems, detect problems in a design idea, and incorporate feedback into a design idea. The difference between the pre and post self-report scores provides an indication of how participants' beliefs change. In the ad design task, across all participants, self-efficacy rose from 10.8 to 12.1 (out of 20); a paired T-tests shows this is significant ($t(31)=2.243$, $p=0.032$). An overall increase in self-efficacy also occurred in the egg drop study, as scores moved from 13.0 to 14.0 ($t(33)=2.321$, $p=0.027$). These increases are consistent with prior findings that show people's self-efficacy beliefs increase after practice [13,31].

In both studies, an analysis of variances was performed with condition (Serial/Parallel) and prior task experience (Experienced/Novice). In ad design, participants in the Parallel condition reported a significant increase in self-efficacy scores, a net gain of 2.5 points ($F(1,28)=4.210$, $p=0.049$), while the Serial condition essentially remained even. However, in egg drop design, condition had no overall effect on the net change in self-efficacy scores ($F(1,30)=0.119$, $p=0.733$). These results align with the different performance findings of the two studies.

Participants with prior ad design experienced reported a similar gain in self-efficacy as novices ($F(1,28)=0.075$, $p=0.787$). In egg-drop design, however, novices reported a net increase in self-efficacy scores (1.9), while those with prior egg drop exposure stayed even ($F(1,30)=4.562$, $p=0.041$). Both studies revealed an interaction effect between condition and prior experience (see Figure 8). Novices in ad design reported a 2.9 increase in self-efficacy in Parallel, but a slight decrease when in Serial (-0.73) ($F(1,28)=6.331$, $p=0.018$). In egg drop design, the opposite interaction effect occurred. Novices reported a 3.9 increase in self-efficacy in Serial, but only a slight increase in Paral-

lel (0.9) ($F(1,30)=4.259$, $p=0.049$). In lay terms, Parallel prototyping had an overall effect on an individual's belief in their ad design ability; this was especially true for novices. First-time egg-drop participants gained the most self-efficacy when placed in the Serial prototyping condition.

DISCUSSION

In the ad design study, the manipulation of Serial vs. Parallel led to creations with better performance by every measure: higher subjective ratings, more impressions served up by MySpace, better click-through rates, more visitors to the client Web site, and more site interaction per visitor. Participants created the same number of prototypes and received the same amount of feedback during an equivalent time period.

Why did the process manipulation impact how participants performed? Design performance improves through an astute discovery of contextual constraints, malleable variables, and their interrelations. There are millions of ways to combine text, images, and backgrounds in a 160×600 pixel ad design. The study showed Parallel participants created significantly more divergent prototypes and began to understand how the moving pieces relate. Parallel feedback instigates inductive reasoning on a set of rival observations and leads to principled choices for subsequent prototypes. In Serial prototyping, participants created similar designs, as ideas tended to follow directly from the feedback ("this piece is wrong; now fix it."). This incremental approach may implicitly encourage designers to hone in on certain variables, while others remain hidden.

A motivational account says ad design performance degrades because Serial participants perceive the expert feedback more critically. The ad study showed Serial participants reported more frustration about the expert feedback. Parallel participants experienced no frustration because they have alternatives to assuage any emotional reaction to criticism. Serial participants could spend their time sensibly analyzing the design space, but instead they are discouraged and lose confidence. As the study demonstrated, Parallel participants gained self-efficacy for the ad design task, while the Serial participants reported no change. As another data point, participants were asked to leave their email if they wanted to volunteer later on for *Ambidextrous* magazine. Twelve out of sixteen Parallel participants provided their email, while only five of seventeen did the same in Serial ($\chi^2=3.32$, $p>0.1$), which suggests the Parallel process helped motivate future action.

Both the cognitive and motivational accounts likely contribute to the overall performance differences, but future work is required to tease apart the relative effects. It raises a question about the causal relationship: did self-efficacy increase because participants perceived improvement, or did performance increase because participants gained self-efficacy through the process? It's probably both.

Why did the Parallel approach affect the outcome for the ad design task, but not for the egg drop? There are several potential explanations. First, with egg drop vessels, partici-

pants receive objective feedback—how the vessel falls and whether the egg breaks—closely linking the interim metric to the ultimate metric. In ad design, participants receive expert critique on multiple variables (e.g., theme, layout, readability...) that are loosely tied to the final performance metrics (e.g., click-through data, client ratings). Ad design critique requires a greater degree of interpretation, which can be misconstrued.

Second, due ephemeral nature of egg-drop trials, Parallel participants cannot directly contrast feedback side-by-side. They can only perceive one test drop at a time, making it difficult to compare the relative efficacy of specific features. In the ad study, Parallel participant had time to study and compare the feedback and then draw conclusions on where to go.

Third, the egg-drop task seemed more sensitive to small variations in implementation, placing a premium on craft. For example, a loose piece of tape can destroy performance. More than half of the egg-drop participants actually achieved a better performance on a prototype, not the final design (18 of 35). In ad design, minor implementation flaws—such as slightly misaligned text—likely do not significantly impact outcome.

The differences between the ad design and egg-drop design tasks shed light on when and why parallel prototyping affects design outcome. The independent variable did not affect egg-drop participants as it did in ad design. The egg-drop feedback did not explicitly support comparison in Parallel, nor did it discourage Serial participants the way it did in ad design.

What are the tradeoffs of using advertisement design as an experimental paradigm? Other digital tasks were considered for this design study, such as Web site design and programming tasks. Ad design provides a useful experimental paradigm because it is brief enough to enable iterative creation and feedback within a single lab session. It also has a rich objective and subjective performance metrics. However, several issues did emerge. The click-through data are difficult to analyze statistically because the number of impressions change based on the prior performance of the ad. A more straightforward setup would be to hold the number of campaign impressions constant across ads, allowing statistical analysis on the number of clicks. The analysis could be further enriched with more information about who clicked on the ads, how many ads users viewed prior to clicking an ad, and how many of the total clicks were repeat-clicks from the same user.

The expert feedback system developed for this study has not been validated. While evidence shows Parallel participants outperformed Serial, it does not demonstrate the feedback actually produced overall better ads than no feedback at all. For the sake of this experiment, however, it does not matter if the ad feedback is intrinsically good or bad; the relative performances demonstrated the effect of the process manipulation.

CONCLUSIONS

This paper investigated the relative effects of parallel and serial prototyping on design performance, learning, and self-efficacy. In a Web advertisement design task, participants in the Parallel condition outperformed Serial participants and reported a significantly higher increase in self-efficacy. Parallel participants created significantly more divergent prototypes and directly compared expert critiques, allowing for a broader exploration of key variables and their interrelations. Notably, Serial participants received critique on each ad design sequentially, which tended to elicit negative reactions. In a mechanical egg-drop design task, the Parallel versus Serial manipulation had no effect on either performance or self-efficacy. The main benefits of parallel prototyping are diluted by the objective and ephemeral nature of the feedback. Future work will focus on unpacking the relative learning and motivational effects of parallel prototyping.

ACKNOWLEDGEMENTS

We would like to thank Ugochi Acholonu, Lera Boroditsky, Daniel Schwartz, Ewart Thomas, and Terry Winograd.

REFERENCES

1. Egg drop contest. http://en.wikipedia.org/wiki/Egg_drop_competition.
2. Google Analytics. <http://www.google.com/analytics/>.
3. Ambidextrous Magazine. <http://ambidextrousmag.org/>.
4. Amazon Mechanical Turk. <https://www.mturk.com/mturk/>.
5. Alexander, C., Ishikawa, S., and Silverstein, M. *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, 1977.
6. Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. Large Stakes and Big mistakes. *Federal Reserve Bank of Boston*, 2005.
7. Arkes, H.R. and Blumer, C. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes* 35, 1 (1985), 124-140.
8. Arora, J. *Introduction to Optimum Design*. Academic Press, 2004.
9. Atman, C.J. and Bursic, K.M. Teaching engineering design: Can reading a textbook make a difference? *Research in Engineering Design* 8, 4 (1996), 240-250.
10. Atman, C.J., Chimka, J.R., Bursic, K.M., and Nachtman, H.L. A comparison of freshman and senior engineering design processes. *Design Studies* 20, 2 (1999), 131-152.
11. Ball, L.J., Evans, J.S.B.T., and Dennis, I. Cognitive processes in engineering design: a longitudinal study. *Ergonomics* 37, 11 (1994), 1753.
12. Ball, L.J. and Ormerod, T.C. Structured and opportunistic processing in design: a critical discussion. *Int. J. Hum.-Comput. Stud.* 43, 1 (1995), 131-151.
13. Bandura, A. *Self-Efficacy: The Exercise of Control*. Worth Publishers, 1997.
14. Boroditsky, L. Comparison and the development of knowledge. *Cognition* 102, 1 (2007), 118-128.
15. Buxton, B. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann, 2007.
16. Carson, S.H., Peterson, J.B., and Higgins, D.M. Reliability, Validity, and Factor Structure of the Creative Achievement Questionnaire. *Creativity Research Journal* 17, 1 (2005), 37-50.
17. Colhoun, J., Gentner, D., and Loewenstein, J. Learning abstract principles through principle-case comparison. *Proceedings of Cognitive Science Society*, (2008), 1659-1664.
18. Cross, N. Expertise in design: an overview. *Design Studies* 25, 5 (2004), 427-441.
19. Cross, N. *Designerly Ways of Knowing*. Springer, 2006.
20. Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*. Harper Perennial, 1991.
21. Dodgson, P. and Wood, J. Self-esteem and the cognitive accessibility of strengths and weaknesses after failure. *Journal of Personality and Social Psychology* 75, 1 (1998), 178-197.
22. Dorst, K. and Cross, N. Creativity in the design process: co-evolution of problem-solution. *Design Studies* 22, 5 (2001), 425-437.
23. Dow, S.P., Heddleston, K., and Klemmer, S.R. The Efficacy of Prototyping Under Time Constraint. *Creativity and Cognition*, (2009).
24. Duncker, K. *On Problem-solving*. Greenwood Press Reprint, 1972.
25. Dweck, C., Mangels, J.A., and Good, C. Motivational effects of attention, cognition and performance. In *Motivation, Emotion, and Cognition: Integrated Perspectives on Intellectual Functioning*. Erlbaum, 2004, 41-55.
26. Dym, C.L. and Little, P. *Engineering Design: A Project-Based Introduction*. Wiley, 1999.
27. Fredrickson, B.L. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist* 56, 3, 218-226.
28. Gick, M.L. and Holyoak, K.J. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1-38.
29. Goel, V. and Pirolli, P. The structure of design problem spaces. *Cognitive Science* 16, 3, 395-429.
30. Gopnik, A., Meltzoff, A.N., and Kuhl, P.K. *The Scientist in the Crib: What Early Learning Tells Us About the Mind*. Harper Paperbacks, 2001.
31. Hall, T.S. Improving Self-Efficacy in Problem Solving: Learning from Errors and Feedback. 2008.
32. Hartmann, B., Klemmer, S.R., Bernstein, M., et al. Reflective physical prototyping through integrated design, test, and analysis. *Proceedings of the 19th annual ACM symposium on User interface software and technology*, ACM (2006), 299-308.
33. Hartmann, B., Yu, L., Allison, A., Yang, Y., and Klemmer, S.R. Design as exploration: creating interface alternatives through parallel authoring and runtime tuning. *Proceedings of the 21st annual ACM symposium on User interface software and technology*, ACM (2008), 91-100.
34. Jansson, D. and Smith, S. Design Fixation. *Design Studies* 12, 1 (1991), 3-11.
35. Kelley, T. *The Art of Innovation*. Profile Business, 2002.
36. Kolodner, J. *Case-Based Reasoning*. Morgan Kaufmann, 1993.
37. Kolodner, J.L. and Wills, L.M. Powers of observation in creative design. *Design Studies* 17, 4 (1996), 385-416.
38. Kosara, R. Visualization Criticism - The Missing Link Be-

tween Information Visualization and Art. *Proceedings of the 11th International Conference Information Visualization*, IEEE Computer Society (2007), 631-636.

39. Laseau, P. *Graphic Thinking for Architects and Designers, 2nd Edition*. John Wiley & Sons Inc, 1988.
40. Lepper, M.R., Master, A., and Yow, W.Q. Intrinsic Motivation in Education. In *Advances in Motivation in Education*. 2008.
41. Mele, A.R. *Motivation and Agency*. Oxford University Press US, 2005.
42. Nickerson, R.S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, (1998), 175--220.
43. Ormrod, J. *Human learning*. Pearson, 2008.
44. Piaget, J. *The Psychology of Intelligence*. Routledge, 2001.
45. Pink, D. The Surprising Science of Motivation. 2009.
46. Rittel, H. and Webber, M. Dilemmas in a general theory of planning. *Policy Sciences* 4, 2 (1973), 169, 155.
47. Schon, D.A. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass, 1990.
48. Schon, D.A. *The Reflective Practitioner: How Professionals Think in Action*. Ashgate Publishing, 1995.
49. Schrage, M. *Serious Play: How the World's Best Companies Simulate to Innovate*. Harvard Business School Press, 1999.
50. Smith, S., Kohn, N., and Shah, J. What You See Is What You Get: Effects of Provocative Stimuli on Creative Invention. *NSF International Workshop on Studying Design Creativity*, (2008).
51. Thompson, L., Gentner, D., and Loewenstein, J. Avoiding Missed Opportunities in Managerial Life: Analogical Training More Powerful Than Individual Case Training. *Organizational Behavior and Human Decision Processes* 82, 1 (2000), 60-75.
52. Tohidi, M., Buxton, W., Baecker, R., and Sellen, A. Getting the right design and the design right. *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM (2006), 1243-1252.
53. Tversky, B., Suwa, M., Agrawala, M., et al. Sketches for Design and Design of Sketches. 2008.

APPENDIX A: Expert Critique Statements

1. Overall/ Thematic

Ambidextrous seeks an ad with a single clear message that matches the theme of their journal.

Ambidextrous wants an ad that clarifies the product: a journal about design and design process.

Ambidextrous desires an ad that is simple, readable, consistent, and deliberate.

Ambidextrous does not want the ad to sound exclusive; they are open to anyone with interest.

Ambidextrous is looking for a creative and clever ad.

Ambidextrous is looking for a professional and tasteful ad.

Ambidextrous wants an exciting and visually appealing ad.

Ambidextrous wants an ad that matches the journal's style.

Ambidextrous wants an ad that reaches out to design practitioners, students, and researchers.

Use graphics/images that support the overall message. What message are you trying to convey?

Use colors/fonts that support the overall message. What message are you trying to convey?

Remember that the ad is a link; the URL does not necessarily have to be on the ad design.

2. Composition & Layout

Visual Flow and Balance

Try to create a balanced layout where the graphics don't tilt to one side or the other.

Try to create a visual flow for the viewer—what should the viewer see first?

Think about the proximity of different elements. How close together or far apart elements are placed suggests a relationship (or lack thereof) between otherwise disparate parts.

To help balance the ad, leave slightly more space at the bottom relative to the top of the ad.

Contrast the position of elements to draw the viewer's attention to the most important parts.

To create consistency for the viewer, create a consistent and balanced look using repetition.

Spacing and Alignment

Align text and graphics to create more interesting, dynamic, and appropriate layouts.

Use alignment to create a clean and organized look.

It's ok to break alignment only to draw the viewer's attention to important elements in the ad.

Use white around text and images to help frame the content.

Use space—the absence of text and graphics—to provide visual breathing room for the eye.

Try to balance the spacing around the border of the ad design. These visual elements in the ad don't line up.

Consider playing around with different ways to justify the text (e.g., center, left, or right-justified).

Emphasis & Hierarchy

Be conscious of competing elements in the ad. Think about what should have emphasis.

Draw the viewer's attention to elements by contrasting size (scale).

Think about the visual hierarchy of the different elements (texts, images, colors, etc) of the ad. What is the most important?

Help the viewer recognize, identify and comprehend the most important information in the ad.

Use elements with visual intensity or color for emphasis.

3. Fonts, Colors, Images

Font Type

Try not to distort the font so that it becomes hard to read.

Use large, bold font/graphics to create focus or emphasis on the ad design.

If using text over an image, make the text bigger and darker than normal; make sure it is readable.

For text to stand out it has to be substantially different than other text.

Try not to mix serif and sans serif fonts.

Avoid using two different fonts that are too similar.

Try not to over-emphasize text elements. (ex. a font does not need to be large, bold, and italic).

Images

Use large, bold graphics to create the focus of the ad design.

Consider using images for more visual impact.

Consider using fewer images.

Try not to over-rotate images, as it often distorts the content.

Color

Use color to create emphasis, to separate different elements, or to categorize content.

Avoid really light, bright colors.

Avoid colors together that look too similar (ex. brown & grey).

Try to use different colors that go well together.

Avoid complicated backgrounds.

Try to create a good visual separation between the text and the background