“I Know It When I See It”: How Experts and Novices Recognize Good Design

Kesler Tanner and James Landay

Abstract Design novices have limited design experience and typically lack the skills or confidence to create good design, however, they may be able to recognize good design. To assess this ability, 53 novice designers and 52 expert designers participated in an online study where they evaluated a series of websites based on aesthetic appeal using two different modes of comparison. Results show that both experts and novices are able to recognize good design and that novices are able to do so almost as well as experts (76.5% accuracy compared to 81.2%). The greatest determinant of whether a participant would correctly identify a higher-rated design was the difference in the two websites’ ground-truth aesthetic ratings. However, expertise and the mode by which the comparison was presented had a significant impact on accuracy (*Keep-the-Best* = 83.6% and *Tournament* = 74.1%).

1 Introduction

While not everyone may be capable of preparing a well-cooked steak, most people may feel they can identify a great steak when eating it, especially if compared with a steak from their local all-you-can-eat buffet restaurant. People may feel similarly about music. Although they may be incapable of composing the next great symphony, their ears can discern between works of great musicians like Mozart and those of the high school rock band practicing down the street.

At some point, however, people’s ability to distinguish between good and bad, whether that be food, music or something else, breaks down. The difference in quality becomes too small to be perceived, and people resort to guessing. Experts through training and experience may develop an increased ability to discern and judge quality such that they can still separate items even when they are indistinguishable to a novice. The question remains, however, how big is that difference? At what point does a person’s ability to assess difference in quality break down?

K. Tanner (✉) · J. Landay
Stanford University, Stanford, CA, USA
e-mail: keslert@stanford.edu; landay@stanford.edu

C. Meinel, L. Leifer (eds.), *Design Thinking Research*, Understanding Innovation,
https://doi.org/10.1007/978-3-319-97082-0_13

There is also a question of subjectivity. Is there a meaningful scale for judging steak quality, or do people's preferences differ too greatly? How subjective are these orderings?

Like food and music preferences, design is an area in which quality is believed to be highly subjective. Design is also similarly pervasive. Millions of people engage in the design process on a daily basis, creating slide presentations, social media graphics, websites, etc. Some of these people are design experts, trained in best design practices with hundreds of hours of study and experience under their belts, but they are the exception. Most people have limited expertise and understanding of design principles.

Even though most people are not actively participating in the design process, they are exposed to numerous examples of design as they browse the internet, go shopping, and drive on the highway. We hypothesize that, similar to a person's ability to recognize high quality food, the constant exposure people have to design causes them to subconsciously create a basic framework within which to judge design. Even though they are not purposefully training to become experts in design, they can use this framework to recognize good design when they see it. We further hypothesize that while good design is subjective, there is a high level of agreement between people in assessment of design quality.

To test these hypotheses, we built on a study conducted by Reinecke and Gajos (2014) in which they explored design preferences of novices throughout the world by having participants complete an online survey where they rated the visual appeal of websites on a scale of 1–9. We conducted a similar survey in which participants rated a selection of the same websites on a scale of 1–9. Additionally, to rate a second selection of Reinecke and Gajos websites, we used two different comparison methods in which the participants were presented with pairs of websites and asked to select which of the two websites was more visually appealing. While design includes more than visual aesthetic appeal, we chose to focus on this aspect of design since visual appeal has a strong correlation on the perceived usability of something (Hassenzahl 2004), and people that find a design appealing are more tolerant of usability issues (Kurosu and Kashimura 1995). We also collected sufficient demographic information from our participants to separate them into novice and expert categories based on their design expertise.

Based on the analysis of our data, we found that novices are almost as accurate as experts when discerning between the aesthetic qualities of two websites. We also found that while design is indeed subjective, there is also a high degree of agreement about the quality of a visual design. Finally, we found that the mechanism used to present design comparisons has an impact on the overall accuracy and time taken.

The main contributions of this work are: (1) a quantitative assessment showing that visual design quality is not purely subjective due to the high degree of agreement, and (2) empirical data showing that novices are almost as accurate at recognizing good design as experts (within 5%). Because of these primary contributions, this research informs better design tools, and the creation of a meaningful design scale that can be created from more easily obtainable novice comparisons.

2 Related Work

In this section we discuss related work regarding differences between experts and novices, obtaining design feedback from novices, and design intuition.

2.1 Differences Between Experts and Novices

Novices and experts by definition are separated based on their level of expertise. With their higher level of expertise, experts are more familiar with successful design practices and principles. They understand the benefits of parallel prototyping and starting broad with many different ideas before honing in on a final solution, compared to novices who tend to take a “depth first” approach, exploring one design at a time (Cross 2004). As demonstrated by Christiaans and Dorst in their study of junior vs. senior industrial design students, even novices who do recognize the need for seeking outside inspiration and exploring multiple ideas tend to get caught in the information gathering stage and are unable to progress to synthesis (Christiaans and Dorst 1992). On the other hand, experts, with time, develop a repository of solutions. When facing a new problem, they map existing solutions [e.g. design patterns (Duyne et al. 2007; Tidwell 2005)] to new problems in creative ways, whereas novices lack this knowledge of existing solutions and attempt to create a new solution for each new problem (Lloyd and Scott 1994).

While past research has focused primarily on distinguishing the creative abilities of novice as compared to expert designers, we seek to expand this line of research to explore the differences (or lack thereof) between the design quality recognition abilities of these two groups.

2.2 Design Feedback from Novices

Due to the difficulty of obtaining feedback from experts, novices have increasingly been turned to for design feedback. This feedback has taken the form of online task workers using an interface, social media requests, or classroom peers writing a critique. Novice feedback has been found to be helpful to designers, but is perceived to be not as valuable as expert feedback. Research has shown that providing novices with a structure or “scaffolding” with which to provide their feedback helps close the gap between the quality of feedback given by experts and novices (Willett et al. 2012; Xu and Bailey 2012). In fact, Alvin Yuan et al. determined that although an online crowd may seem to lack relevant domain experience, by requiring a non-expert crowd to use a rubric to provide feedback, novice feedback was “rated nearly as valuable as expert feedback” (Yuan et al. 2016). Furthermore, this gap becomes even less significant when timeliness and “clear messag[ing] to a target audience” are the primary concerns of the needed feedback, as opposed to a “range and depth of feedback” (Xu et al. 2015).

Systems such as Voyant (Xu et al. 2014) and CrowdCrit (Luther et al. 2014) demonstrate the validity, effectiveness, and value of design feedback from a non-expert crowd. Such systems are able to eliminate the need to expend social capital to obtain peer critique and the feedback obtained was also determined by experts to approach the quality of peer critique that could be “enthusiastic[ally]” incorporated into design improvements (Luther et al. 2015). While this research clearly shows that novices have the ability to provide meaningful feedback, we sought to delve deeper to validate the underlying assumptions used in this research, including whether novices are able to recognize good design on par with experts and if there is agreement as to what constitutes *good* design.

2.3 Design Intuition

Experts’ intuitive design abilities have been the subject of a large body of research. This research breaks down experts’ power of intuition into two primary functions: generating alternatives (intuitive speculation) and choosing between these alternatives (intuitive impulse) (Faste 2017). Emphasis is placed on the learned nature of this intuition (Faste 1995; Petitmengin-peugeot 1999), with Cross claiming that this intuition is “honed over time, the ability to make these sorts of qualitative decisions can be considered the designer’s systemic (‘intuitive’) method, through which insight and technical mastery are developed” (Cross 2004). While experts are certainly actively honing their intuitive abilities, we believe that novices may possess a similar intuitive impulse, or the ability to recognize good design without the purposeful honing of this ability.

3 Experiment

We conducted a study to evaluate the effect of design expertise (novice and expert) and comparison mode (*Keep-the-Best* and *Tournament*) on a person’s ability to recognize good design. Our goal was to discover to what degree people generally agree upon a website’s aesthetic appeal and how novices and experts differed in their perceptions.

3.1 Materials

We used the website snapshots from Reinecke’s data set (Reinecke and Gajos 2014), excluding foreign websites and some we felt were overly recognizable (e.g., Boy Scouts of America and Disney World). This resulted in a total of 338 websites. These websites were originally selected by Reinecke to represent a range of colorfulness, visual complexity, and genre.

3.2 Participants

Two hundred and six participants took part in the study, and were found using convenience sampling from sources such as Slack designer communities, NextDoor, Reddit, and Facebook. People were not compensated for their participation, but were told they would be shown how they compared with others upon the study’s completion. We then filtered participants to include only those who completed the study within a reasonable time (2 hours), took the survey from within the United States, did not experience any technical difficulties, and professed to have completed the survey to the best of their ability. After this filtering, 118 participants were remaining.

Participants were then separated into three possible groups based on the following two questions: (1) “Do you or have you worked as a design professional?” and (2) “How many years have you worked as a design expert?”

If a participant answered “*No”* to question 1, they were placed in the *Novice* group. If a participant answered “*Yes”* to question 1 and claimed to have 2 or more years of experience, they were placed in the *Expert* group. If a participant answered “*Yes”* to question 1 and claimed to have less than 2 years of experience they were not included in the study. This left us with 53 novice and 52 expert participants. Of these participants, 50 identified as female and 55 identified as male, and the average age was 32.5 years (SD = 11.3, MIN = 19, MAX = 74). For those classified as experts, the average years worked as a design professional was 5.5 (SD = 3.2).

3.3 Apparatus

The study was conducted on s*******.com, a platform we built to conduct design studies. The website and studies were built using React.[1] The server, including hosting and database, was built using Firebase.[2] Images used in the studies were pre-fetched at the start of the study to ensure no delay occurred during the actual

[1] https://facebook.github.io/react/

[2] https://firebase.google.com/

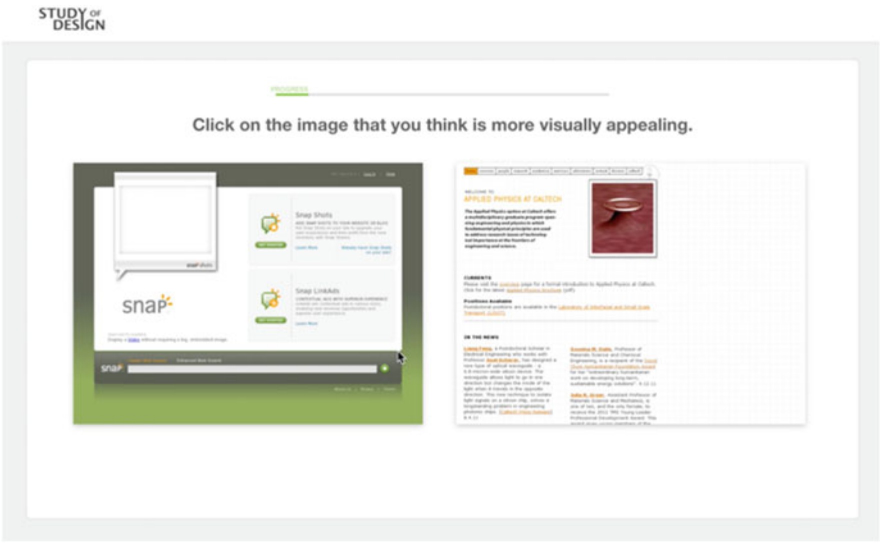
study. Participants took part in the study remotely, using a personal desktop or laptop. The study could not be taken on mobile devices.

3.4 Procedure

Upon arriving at s*******.com, participants began the study by completing a short set of demographic questions (as used for filtering described above). The main part of the study consisted of three tasks: *Rating*, *Keep-the-Best*, and *Tournament*. The order in which a participant completed these tasks was randomized, and the set of websites used in each task were unique to the task.

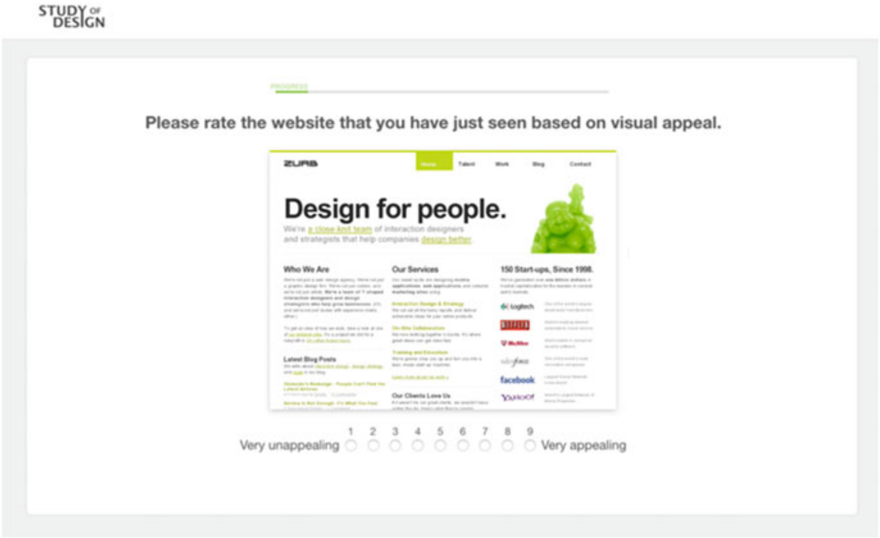
During the study, a participant viewed 128 distinct websites (*Rating* = 64, *Keep-the-Best* = 32, *Tournament* = 32). To compile the set of images for each task, we used the 1–9 ratings collected from Reinecke's study to provide an average score for each website. We used this score to order the websites from highest rated to lowest rated. We then divided this spectrum into 16 equally sized buckets. From each of the 16 buckets we drew a random image. This process was repeated twice for the *Keep-the-Best* and *Tournament* tasks, and four times for the *Rating* task.

During the *Rating* task (see Fig. 1), users were asked to rate a website based on its aesthetic appeal from 1 (very unappealing) to 9 (very appealing). A website was shown for 500 milliseconds after which it would disappear and the participant would provide a score. In total each participant rated 64 distinct websites.



Click on the image that you think is more visually appealing.

Fig. 1 Interface used during *Rating* task. Participants rated a website on a scale from 1 to 9 to advance to the next decision



Please rate the website that you have just seen based on visual appeal.

1 2 3 4 5 6 7 8 9
Very unappealing ○ ○ ○ ○ ○ ○ ○ ○ ○ Very appealing

Fig. 2 Interface used during the *Keep-the-Best* and *Tournament* tasks. Participants clicked on the more appealing website to advance to the next decision

The *Keep-the-Best* and *Tournament* comparison tasks were similar. Participants were presented with two websites and asked to select the website that was more visually appealing (see Fig. 2). In the *Keep-the-Best* task, the selected website would be included in the immediate next round and compared against a new website. In the *Tournament* task, the selected website would be added to a “winners pool” and would resurface in future comparisons when all other websites from the current cycle had been compared. Each comparison task included 32 websites, resulting in 31 comparisons for each.

Before beginning each task, the participant was provided with directions and a brief training set using the same four websites (these did not appear in any of the participant’s own comparisons or ratings). At the end of each task the participant filled out a NASA TLX form to assess perceived workload. Between each task, participants were given a break (as long as desired) before continuing on to the next section.

At the end of the online survey, participants were asked if they experienced any technical errors, and if they had completed the survey to the best of their ability. On average, the survey took participants 10–15 minutes.

3.5 Data Preparation and Analyses

Our final dataset consisted of 6720 website ratings (3328 expert ratings) and 6510 website comparisons (3224 expert comparisons).

To assess whether a decision was correct during the *Keep-the-Best* and *Tournament* tasks, we needed to establish a system of ground-truth values for the websites. This was originally done using an aggregate of the ratings collected by Reinecke. As an alternative system, we used the data collected during the Rating task, which consisted of 3392 novice and 3328 expert ratings. We found that using the combined expert and novice ratings from our study produced a higher accuracy for both novices and experts during the *Keep-the-Best* and *Tournament* tasks. Since the websites a participant saw in the *Rating* task were unique to that task, a participant's ratings did not improve their own accuracy. Comparing the ordering of websites from Reinecke's and Gajos' data against those obtained in our study yielded a Pearson correlation score of 0.74. We hypothesize that the difference between the two orderings can partially be explained by a difference in 4 years of being collected and the increased percentage of experts in our study.

For each comparison between two websites made during the *Tournament* and *Keep-the-Best* tasks, the following metrics were calculated:

Time: the time in milliseconds from the moment the image appeared to the moment the participant clicked on an image.

Correct: a decision was marked as correct if the participant clicked on the website image that had a higher ground-truth score.

Absolute Difference: the absolute difference between the two websites' scores being compared.

4 Results

Our analysis of variance showed that mode order did not exhibit a main effect. We present our results as a function of *Expertise*, *Mode* and *Absolute Difference*.

4.1 Accuracy

We ran a generalized linear mixed model fit by maximum likelihood to examine the main effects and interaction effects of Figs. 3 and 4, measuring accuracy as a function of *Expertise*, *Mode*, and *Absolute Difference*. There was a significant main effect of *Absolute Difference* on accuracy ($\chi 2$ (1,N = 6510) = 545.04, $p < 0.0001$), as larger absolute differences between websites caused increased accuracy. There was no significant main effect of *Expertise* or *Mode.*

There was an interaction effect between *Absolute Difference* and *Expertise* ($\chi 2$ (1, N = 6510) = 18.68, $p < 0.0001$) as design expertise caused increased accuracy at certain levels of *Absolute Difference.* An expert participant had an average accuracy of 63.4% when the absolute difference between two websites was 0.5 compared to an average accuracy of 62.8% of a novice for the same type of comparison. This

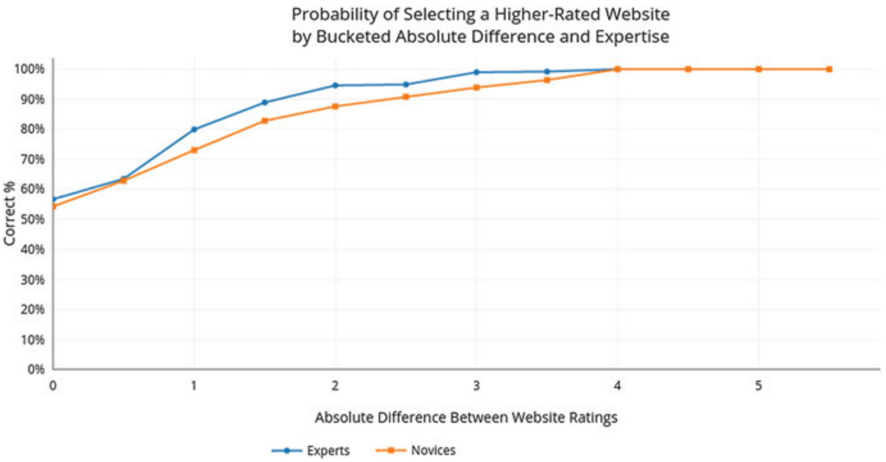


Fig. 3 Probability of selecting a higher-rated website as a function of Absolute Difference and Expertise

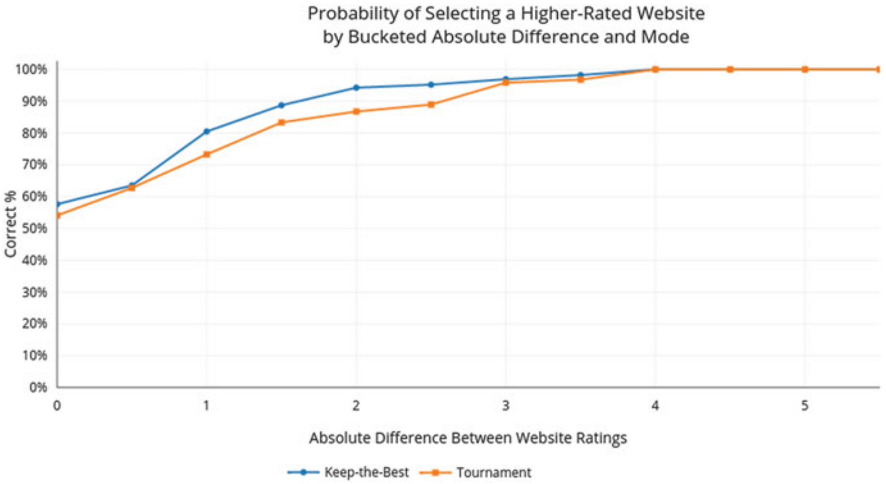


Fig. 4 Probability of selecting a higher-rated website as a function of Absolute Difference and Mode

accuracy increased to 88.9% for an expert and 82.8% for a novice when the absolute difference was ~1.5. When the absolute difference was 4.0 or greater, both experts and novices were 100% accurate. Additional details can be seen in Table 1.

An interaction effect also existed between *Absolute Difference* and *Mode* ($\chi 2$ (1, N = 6510) = 6.91, $p < 0.01$) as the comparison mechanism caused increased accuracy at certain levels of *Absolute Difference.* When comparing two websites during the *Keep-the-Best* task, participants had an average accuracy of 63.5% when the absolute difference between those two websites was ~0.5 compared to an

Table 1 Top shows average accuracy and number of decisions as a function of Expertise and Bucketed Absolute Difference. Bottom shows average accuracy and number of decisions as a function of Mode and Bucketed Absolute Difference

	Bucketed Absolute Difference											
	0	0.5	1*	1.5*	2*	2.5	3*	3.5	4	4.5	5	5.5
Expert												
Accuracy	56.6%	63.4%	79.9%	88.9%	94.6%	94.9%	99.0%	99.2%	100%	100%	100%	NA
# Decisions	362	644	571	530	387	316	198	119	66	29	2	0
Novice												
Accuracy	54.2%	62.8%	73.0%	82.8%	87.6%	90.8%	93.9%	96.4%	100%	100%	100%	100%
# Decisions	371	683	588	512	450	314	180	111	49	17	9	2
	Bucketed Absolute Difference											
	0	0.5	1**	1.5**	2**	2.5	3	3.5	4	4.5	5	5.5
Keep-the-Best												
Accuracy	57.6%	63.5%	80.5%	88.7%	94.2%	95.2%	96.9%	98.2%	100%	100%	100%	100%
# Decisions	269	581	497	496	452	395	258	168	87	41	10	1
Tournament												
Accuracy	54.1%	62.7%	73.3%	83.3%	86.8%	88.9%	95.8%	98.6%	100%	100%	100%	100%
# Decisions	464	662	588	546	385	235	120	62	28	5	1	1

Results where $p < 0.001$ are marked with ** and results where $p < 0.05$ are marked with asterisks average accuracy of 62.7% during the *Tournament* task. This accuracy increased to 88.7% in the *Keep-the-Best* task and 83.3% during the *Tournament* task when the absolute difference was ~1.5. When the absolute difference was 4.0 or greater, participants were 100% accurate during both the *Keep-the-Best* and *Tournament* tasks. Additional details can be seen in Table 1.

Figures 3 and 4 illustrate the connection between absolute difference and accuracy. When the absolute difference between two websites' scores approaches zero, the probability of choosing the higher-rated website approaches 50%. As the absolute difference between two websites' scores increases, the probability of choosing the higher-rated website also increases until it reaches a maximum accuracy of 100%. Table 1 shows that this point of perfect accuracy occurs when the difference between two websites is 4.0.

There was no significant interaction effect between Mode and Expertise, or *Absolute Difference, Mode, and Expertise.*

4.2 Decision Time

We ran a generalized linear mixed effect model to examine main effects, measuring decision time as a function of *Expertise* and *Mode*. In this model, we included the *Rating* task data. There was a significant main effect of *Mode* on decision time ($\chi2$ (2,N = 13,231) = 98.17, $p < 0.0001$), as decisions made during the *Tournament* task took more time than decisions during the *Keep-the-Best* task. The median decision time during the *Tournament* task was 3234 milliseconds. During the *Keep-the-Best* task, the median decision time was 1965 milliseconds. During the *Rating* task, the median decision time was 2417 milliseconds.

There was no significant main effect of *Expertise* on decision time.

4.3 Nasa TLX

Five of the six NASA TLX categories were used to determine a perceived cognitive load for each task: effort, frustration, mental demand, temporal demand, and performance. Each was rated on a 21-point scale where 1 = Very Low and 21 = Very High. For the performance metric, the labels were adapted to 1 = Failure and 21 = Perfect. Perceived cognitive task load was calculated by averaging the individual scores from each category.

We ran a linear mixed effect model to examine a main effect of perceived cognitive load as seen in Fig. 5. There was a significant main effect of *Mode* on cognitive load (F(2206) = 59.86, $p < 0.001$), as the comparison tasks caused a lower perceived cognitive load than a rating task. Both *Tournament* (T(206) = 10.57, $p < 0.001$) and *Keep-the-Best* (T(206) = 16.01, $p < 0.001$) had a perceived lower cognitive load than *Rating*. *Keep-the-Best* also had a perceived lower cognitive load

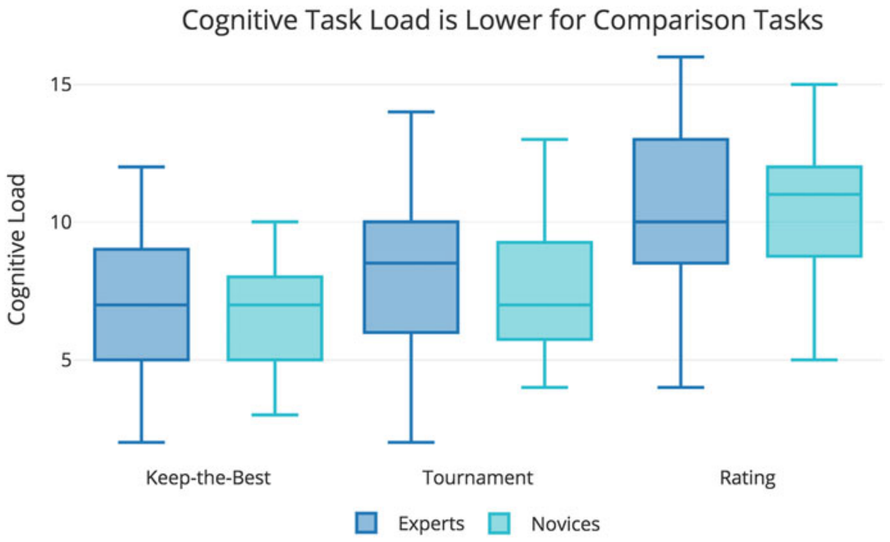


Fig. 5 Perceived cognitive load as self-reported by participants for each of the three tasks

than *Tournament* (T(206) = 5.44, $p < 0.001$). There was no significant effect of *Expertise* on cognitive load.

5 Discussion and Future Work

The goal of this study was to determine if novices are able to accurately recognize good design and how their ability compares to that of an expert. We were able to determine that not only are novices able to recognize good design, but they are able to do so almost as well as experts (within ~5%). While past work (Faste 2017) emphasizes the importance of a learned intuition through the pursuit of professional design experience, our research shows that a participant's natural design intuition is an incredibly powerful ability that can be used without additional training.

5.1 Design Subjectivity and Agreement

The general feeling surrounding design is that its aesthetic quality is purely subjective and cannot be quantified. Our study, however, demonstrated empirically that while design is indeed subjective, there is also a high degree of agreement between both experts and novices, and that some degree of quantification is possible. While our absolute numerical rankings may not be perfect, both they and the comparisons support this correlation. This is further strengthened by the largest signal for whether a study participant would choose a higher-rated website being the absolute difference between those two websites' ratings rather than a participant's level of design expertise or the mode by which the comparison was made. This strengthens the argument that while the aggregated website scores may not be a perfect representation of a website's aesthetic appeal, they are accurate enough to provide a strong indication. While we do agree that design is subjective, it appears that it is highly agreeable and that while a certain design may speak to an individual, on average there is a consensus as to what constitutes good visual design.

5.2 Mode Matters

We found that the *Keep-the-Best* mechanic was superior to the *Tournament* mechanic when asking participants to make comparisons. Not only was it more accurate, but also 50% faster and it required a lower cognitive load. This discovery was particularly surprising based on the anecdotal feedback we received during pilot studies that participants felt significantly more mental baggage during the *Keep-the-Best* task. Participants expressed feeling either that they needed to get rid of designs purely because they'd held on to them for too long, or that they became attached to the design they picked repeatedly and felt they had to continue to choose it to validate past decisions. Numerous participants also cited anxiety regarding how once they eliminated a website during the *Keep-the-Best* task it was gone forever. (Even though websites eliminated during the *Tournament* were also gone forever, no participants cited anxiety relative to this.) The NASA TLX performed as part of the actual study, however, showed that study participants found the *Keep-the-Best* mechanic to be easier than the *Tournament* mechanic, so it is possible that the participants of the pilot study felt more self-conscious with an observer actively watching (and judging) their decisions.

The fact that the *Keep-the-Best* mechanic is not only more accurate and less mentally taxing, but also more closely mimics real-life opportunities (such as parallel prototyping) and applications (including new tools we envision) for comparing designs is promising. That being said, in this study, only two comparison methods were tested. The difference in accuracy, speed, and cognitive load existing between just these two modes suggests that exploring different methods of comparisons could further improve the accuracy, speed, and ease with which websites (or other design tasks) could be examined and compared. Furthermore, based on the interaction effect seen in this study between the absolute difference in ratings between two websites and the mode used, comparison methods could potentially be further refined by incorporating machine learning to optimize the comparison mode based on the absolute difference of the website scores being compared.

5.3 Comparison vs. Rating

We discovered that there were several advantages to comparing two designs and choosing the more visually appealing over rating a design on a 1 to 9 scale. First, participants experienced a lower perceived cognitive load doing a comparison than a rating task. We believe this might be in part because they did not have to maintain a framework for all designs in their minds. To assign a numerical score to a website, a participant needed to have a framework for what constitutes a 3 and how that compares to a 5. Instead, in a comparison, they only needed to compare two websites and determine which was better. The perceived cognitive task load was further decreased in the *Keep-the-Best* mode. This makes sense because in each new round only one new website was introduced, whereas in the *Tournament* mode, a participant saw two new websites each round, essentially doubling the workload.

Second, the median time per decision during the *Keep-the-Best* task was lower than the time per decision during the *Rating* task. This is not substantial when considered as a mere 450 millisecond difference, but collectively over the course of many decisions, it constitutes a 19% speedup. Combined with the lower cognitive load, this type of comparison improves a participant's efficiency while demanding less mental energy.

A third benefit of comparison over rating is it is unnecessary to perform a normalization of user ratings. While one participant might rate websites in the range of 1–6, another might rate the same websites from 4–7. With enough participants, these differences smooth out, however, with comparisons there is no need for a shared framework, as only the order between the two websites matters.

5.4 Future Design Tools

The ability of novice designers to recognize good design to a degree that is on par with design experts is underutilized in current design tools. This idea has surfaced in design tools such as Designscape (O'Donovan et al. 2015), Design Galleries (Marks et al. 1997), and Sketchplore (Todi et al. 2016), but these tools are only tapping the surface of this ability. Often novices start their design task from a blank canvas or a pre-built template. They then iterate on their designs by imagining a change in their mind and then carrying out that change on their canvas.

Instead, we envision tools that focus on a user's ability to recognize good design over their ability to first imagine, and then create it. Such tools could work by rapidly exploring the possibilities in a space and relying on the user as an oracle to validate positive exploration paths. This idea shifts the responsibility of the user from identifying *what might look good*, to identifying *what does look good*. Instead of a user changing the background from white to red, they might instead specify that they want to explore background options. They could then be presented with colors, gradients, images, and patterns that might look good. Although the user only specified background options, such a system should be smart enough to adjust font colors, or text shadows to improve legibility. The user is not required to notice what adjustments are being made, but simply that a particular design is an improvement.

In building these new tools, it is important that differences in proposed alternatives be large enough for comparisons to be valuable. As demonstrated by this study, if designs are too similar in their underlying rating, users will be unable to recognize which is better. In situations where only small, incremental changes are possible, we propose that looping back to an earlier design could confirm that the user is incrementally moving in the right direction, rather than incrementally slipping towards a worse design.

We hypothesize that such tools could also help novices and experts to be less prone to fixation effects (Buxton 2010; Dow and Klemmer 2010), while encouraging good design principles such as parallel prototyping (Dow et al. 2012), iteration (Bogumil 1985; Dow and Klemmer 2010; Hartmann et al. 2006; Salter and Whyte 2004) and seeking external feedback (Tohidi et al. 2006).

5.5 Meaningful Design Scale

Our study demonstrated empirically that while design is indeed subjective, there is also a high degree of agreement between both experts and novices. This finding, combined with the ability of novices as a collective to recognize good design, confirms the validity of the creation of a Meaningful Design Scale (MDS), where any design could be ranked based on aesthetic appeal relative to other previously ranked designs and assigned a numeric score. Creating the MDS would not only provide a framework within which to discuss what is good design in a more quantitative way, but could also provide timely and meaningful feedback to augment the current parallel prototyping feedback loop. This could be accomplished by crowdsourcing a series of novice comparisons to place a design on the scale. While peer feedback asks people similar in ability to assess each other's in-progress work (Kulkarni et al. 2013), our findings suggest that novices could collectively provide useful feedback to experts.

Used in this way, companies could incorporate the MDS as a QA measure for their internal design system (i.e., a landing page can only be put live once it has an MDS score of 8+). We hypothesize that an MDS could be made even more valuable to industry and designers in general by further segmenting novice ratings by demographics (i.e., a women's clothing line could require their landing pages to receive an MDS score of 8+ as assigned by women age 18–34 in the United States). Further research could explore how demographics impact the degree of agreement of perceived aesthetic quality of a design.

6 Conclusion

In this research, we conducted a study to understand how well novices recognize good design. We discovered that novices can recognize good visual design almost on par with experts, and that while design is subjective, there is also a high level of agreement. In addition, we learned that the mode by which a comparison is made has a significant impact on accuracy. We discuss the importance of novice's ability to recognize good design and propose how design tools might better leverage this natural skill, as well as how novices' input could be used to create a meaningful design scale, providing a quantifiable means of discussing design.

Acknowledgments We'd like to thank Jacob O. Wobbrock, Rob Semmens, Kurt Gee and Roy Roy for their assistance with statistics. Any opinions, findings, conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect those of any supporter listed above.

References

Bogumil, R. J. (1985). The reflective practitioner: How professionals think in action. *Proceedings of the IEEE* [Internet]. [cited 2017 Sep 15], *73*(4), 845–846. Available from: http://ieeexplore.ieee.org/document/1457478/

Buxton, B. (2010). *Sketching user experiences: Getting the design right and the right design* [Internet]. [cited 2017 Sep 15]. Available from: https://books.google.com/books?hl=en&lr=&id=2vfPxocmLh0C&oi=fnd&pg=PP1&dq=Sketching+User+Experiences:+Getting+the+Design+Right+and+the+Right+Design&ots=06Kuhjl6VO&sig=5VvNqRGGzjzH0ufiVvL8A8UlZXo

Christiaans, H. H. C. M., & Dorst, K. H. (1992). *Cognitive models in industrial design engineering: A protocol study*. American Society of Mechanical Engineers Design Engineering Division.

Cross, N. (2004). Expertise in design: An overview. *Design Studies*. [Internet]. Elsevier; Sep 1 [cited 2017 Sep 15];*25*(5), 427–441. Available from: http://www.sciencedirect.com/science/article/pii/S0142694X04000316

Dow, S., & Klemmer, S. R. (2010). The efficacy of prototyping under time constraints. *Design Thinking*. [Internet]. [cited 2017 Sep 15], pp. 111–128. Available from: http://dl.acm.org/citation.cfm?id=1640260

Dow, S. P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L., & Klemmer, S. R. (2012). Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *Design Thinking Research Studies*. Co-Creation Practice [Internet]. [cited 2017 Sep 15], pp. 127–153. Available from: http://link.springer.com/chapter/10.1007/978-3-642-21643-5_8

Faste, R. (1995). The role of aesthetics in engineering. *Japan Society of Mechanical Engineering Journal*. [Internet]. [cited 2017 Sep 15]. Available from: http://www.fastefoundation.org/publications/the_role_of_aesthetics.pdf

Faste, H. (2017). Intuition in design: Reflections on the iterative aesthetics of form. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems –* CHI EA '17 [Internet]. [cited 2017 Sep 15], pp. 3403–3413. Available from: http://dl.acm.org/citation.cfm?id=3025534

Hartmann, B., Klemmer, S., Bernstein, M., Abdulla, L., Burr, B., Robinson-Mosher, A., et al. (2006) Reflective physical prototyping through integrated design, test, and analysis. In *Proceedings of 19th Annual ACM Symposium on User interface Software Technology* [Internet]. [cited 2017 Sep 15], pp. 299–308. Available from: http://dl.acm.org/citation.cfm?id=1166300

Hassenzahl, M. (2004, December 1). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interactions*. [Internet]. L. Erlbaum Associates [cited 2017 Sep 15], *19*(4), 319–349. Available from: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci1904_2

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., & Cheng, J., et al. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*. [Internet]. [cited 2017 Sep 15], *20*(6), 1–31. Available from: http://link.springer.com/chapter/10.1007/978-3-319-06823-7_9

Kurosu, M., & Kashimura, K. (1995). Creativity of. Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability. In *Proceedings of ACM Conference on Human Factors in Computer Systems* [Internet]. New York, New York, USA: ACM Press; [cited 2017 Sep 15], 292–293. Available from: http://portal.acm.org/citation.cfm?doid=223355.223680

Lloyd, P., & Scott, P. (1994). Discovering the design problem. *Design Studies*. [Internet]. [cited 2017 Sep 15], *15*(2), 125–140. Available from: http://www.sciencedirect.com/science/article/pii/0142694X94900205

Luther, K., Pavel, A., Wu, W., Tolentino, J., Agrawala, M., & Hartmann, B., et al. (2014) CrowdCrit: Crowdsourcing and aggregating visual design critique. In *Proceedings of Companion Publ. 17th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW Companion ’14* [Internet]. [cited 2017 Sep 15], pp. 21–24. Available from: http://dl.acm.org/citation.cfm?id=2556788

Luther, K., Tolentino, J., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., et al. (2015) Structuring, aggregating, and evaluating crowdsourced design critique. In *Proceedings of 18th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW ’15* [Internet]. [cited 2017 Sep 15], pp. 473–485. Available from: http://dl.acm.org/citation.cfm?id=2675283

Marks, J., Ruml, W., Ryall, K., Seims, J., Shieber, S., & Andalman, B., et al. (1997). Design galleries: a general approach to setting parameters for computer graphics and animation. In *Proceedings of 24th Annual Conference on Computer Graphics Interactive Technology – SIGGRAPH ’97* [Internet]. [cited 2017 Sep 15], pp. 389–400. Available from: http://dl.acm.org/citation.cfm?id=258887

O’Donovan, P., Agarwala, A., & Hertzmann, A. (2015). DesignScape: Design with interactive layout suggestions. In *Proceedings of 33rd Annual Conference on Human Factors in Computer Systems – CHI ’15* [Internet]. [cited 2017 Sep 15], pp. 1221–1224. Available from: http://dl.acm.org/citation.cfm?id=2702149

Petitmengin-peugeot, C. (1999). The intuitive experience. *Journal of Consciousness Studies*. [Internet]. [cited 2017 Sep 15], *2*, 43–77. Available from: http://www.ingentaconnect.com/content/imp/jcs/1999/00000006/f0020002/928

Reinecke, K., & Gajos, K. Z. (2014). Quantifying visual preferences around the world. In *Proceedings of 32rd Annual ACM Conference on Human Factors in Computer Systems – CHI ’14* [Internet]. ACM Press, New York, USA; [cited 2017 Sep 15], pp. 11–20. Available from: http://dl.acm.org/citation.cfm?doid=2556288.2557052

Salter, A., & Whyte, J. (2004). Serious play: How the world’s best companies simulate to innovate [Internet]. *Technovation*. [cited 2017 Sep 15], pp. 277–278. Available from: https://books.google.com/books?hl=en&lr=&id=3f6UdmTaAH0C&oi=fnd&pg=PR9&dq=Serious+Play:+How+the+World’s+Best+Companies+Simulate+to+Innovate&ots=RFdJu21jIC&sig=rHiLvl2c91_MH8Zao_nEZmabVhU

Tidwell, J. (2005). *Designing interfaces: Patterns for effective interaction design* [Internet]. OReilly Media Inc. [cited 2017 Sep 19]. p. 352. Available from: https://books.google.com/books?hl=en&lr=&id=5gvOU9X0fu0C&oi=fnd&pg=PR11&dq=Designing+Interfaces:+Patterns+for+Effective+Interaction+Design&ots=sSZ0N7X9VT&sig=eeDGYaJ91-vKulFUKyqqGI2fhBk

Todi, K., Weir, D., & Oulasvirta, A. (2016). Sketchplore: Sketch and explore with a layout optimiser. In *Proceedings of 2016 ACM Conference* [Internet]. [cited 2017 Sep 15]; Available from: http://dl.acm.org/citation.cfm?id=2901817

Tohidi, M., Buxton, W., Baecker, R., & Sellen, A. (2006). Getting the right design and the design right: Testing many is better than one. In *Proceedings of CHI 2006 Conference on Human Factors on Computer Systems* [Internet]. [cited 2017 Sep 15], pp. 1243–1252. Available from: http://dl.acm.org/citation.cfm?id=1124960

Van Duyne, D. K., Landay, J. A., & Hong, J. I. (2007). *The design of sites: Patterns for creating winning web sites* [Internet]. [cited 2017 Sep 19]. Available from: https://books.google.com/books?hl=en&lr=&id=eE2TxLtDsL8C&oi=fnd&pg=PR13&dq=The+Design+of+Sites:+Patterns+for+Creating+Winning+Web+Sites&ots=v9JzHZPmCj&sig=FwQSMbSU8bI_Nno-Gu0-1Wgh9uc

Willett, W., Heer, J., & Agrawala, M. (2012). Strategies for crowdsourcing social data analysis. In *Proceedings of 2012 ACM Annual Conference on Human Factors on Computer Systems – CHI '12* [Internet]. [cited 2017 Sep 15], p. 227. Available from: http://dl.acm.org/citation.cfm?id=2207709

Xu, A., & Bailey, B. P. (2012) What do you think? A case study of benefit, expectation, and interaction in a large online critique community. In *Proceedings of 15th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW'12* [Internet]. [cited 2017 Sep 15], 295. Available from: http://dl.acm.org/citation.cfm?id=2145252

Xu, A., Huang, S.-W., & Bailey, B. (2014). Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of 17th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW '14* [Internet]. [cited 2017 Sep 15], pp. 1433–1444. Available from: http://dl.acm.org/citation.cfm?id=2531604

Xu, A., Rao, H., Dow, S. P., & Bailey, B. P. (2015). A classroom study of using crowd feedback in the iterative design process. In *Proceedings of 18th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW '15* [Internet]. [cited 2017 Sep 15], pp. 1637–1648. Available from: http://dl.acm.org/citation.cfm?id=2675140

Yuan, A., Luther, K., Krause, M., Vennix, S. I., Dow, S. P., & Hartmann, B. (2016). Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of 19th ACM Conference on Computer Supported Cooperative Work and Social Computing – CSCW '16* [Internet]. [cited 2017 Sep 15], pp. 1003–1015. Available from: http://dl.acm.org/citation.cfm?id=2819953