Combining Crowdsourcing and Learning to Improve Engagement and Performance

Mira Dontcheva
Adobe Research
San Francisco, CA
mirad@adobe.com

Robert Morris
MIT Media Lab
Cambridge, MA
rmorris@media.mit.edu

Joel Brandt
Adobe Research
San Francisco, CA
joel.brandt@adobe.com

Elizabeth M. Gerber
Northwestern University
Evanston, IL
egerber@northwestern.edu

Figure 1. LevelUp for Photoshop is a crowdsourcing platform that combines learning and creative work. Workers learn photo editing skills, while improving real-world images. Interactive step-by-step tutorials teach workers new techniques and Challenge Rounds filled with images from different requester organizations test worker knowledge. The worker interface is shown on the left; several original and improved images are shown on the right.

ABSTRACT

Crowdsourcing complex creative tasks remains difficult, in part because these tasks require skilled workers. Most crowdsourcing platforms do not help workers acquire the skills necessary to accomplish complex creative tasks. In this paper, we describe a platform that combines learning and crowdsourcing to benefit both the workers and the requesters. Workers gain new skills through interactive step-by-step tutorials and test their knowledge by improving real-world images submitted by requesters. In a series of three deployments spanning two years, we varied the design of our platform to enhance the learning experience and improve the quality of the crowd work. We tested our approach in the context of *LevelUp for Photoshop*, which teaches people how to do basic photograph improvement tasks using Adobe Photoshop. We found that by using our system workers gained new skills and produced high-quality edits for requested images, even if they had little prior experience editing images.

Author Keywords

Crowdsourcing; training; games.

ACM Classification Keywords

H.5.2 User Interfaces: Graphical user interfaces

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

CHI 2014, April 26–May 1, 2014, Toronto, ON, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2473-1/14/04$15.00.
http://dx.doi.org/10.1145/2556288.2557217

INTRODUCTION

Many of today's crowdsourcing marketplaces are characterized by cheap labor and simple tasks that are easy for humans and hard for computers, such as tagging images, transcribing text, and ranking search results. For many reasons, platforms like Amazon's Mechanical Turk (MTurk) are ill-suited for complex creative tasks. While researchers have found some ways to remedy this problem [4, 12], crowdsourced creative work remains somewhat elusive, largely because most online workforces lack the expertise or motivation to produce high quality creative output. Some crowdsourcing marketplaces, such as oDesk[1] or 99designs,[2] offer the possibility of finding more advanced workers to accomplish specialized tasks, but specialized skills typically come at a higher cost.

One solution is to simply train the crowdworkers. Unfortunately, most crowdsourcing platforms are not designed to incorporate nuanced and sophisticated tutorials. As a result, it can be hard to help workers master the skills needed for complex creative work. There are a few exceptions to this status quo. Platforms like Samasource[3] and MobileWorks[4] do in-person training to teach crowdworkers computing skills, ranging from basic computer literacy to more advanced skills like web design. These skills are not only useful for the tasks these companies crowdsource, they are also highly valuable for the workers, most of whom come from the developing

[1] odesk.com
[2] 99designs.com
[3] samasource.com
[4] mobileworks.com

world. However, such training is costly, and more automated training tools are needed to help train workers while they engage in crowdsourced labor.

In our work, we help workers develop skills that are both useful for completing work in the crowdsourcing platform and potentially marketable in other contexts. Specifically, we introduce a platform that is both a training tool and a mechanism for producing crowdsourced results. We look specifically at photograph enhancement tasks as a representative form of creative work requiring complex skills. While it is becoming easier to capture high quality photos and apply premade filters, many enhancements are still beyond the reach of fully automated methods. To apply professional-level improvements, an individual must learn how to use many different tools. Moreover, they must gain skills in basic color theory and composition in order to correct white balance, adjust lighting, and crop images aesthetically.

Our training and crowd work platform, called *LevelUp for Photoshop*, is an interactive tutorial game for the popular photo editing software Adobe Photoshop. Through a series of interactive tutorials, LevelUp for Photoshop teaches people how to edit photographs one step at a time. Each interactive tutorial teaches a specific skill, such as how to crop to improve composition. As the workers complete the tutorials, they are encouraged to test their skills by improving real-world images. These images are submitted by requesters who want to crowdsource photo enhancements, such as the Wildlife Center for Silicon Valley, which needs high-quality photographs for the center website and brochures.

To understand the factors that influence learning outcomes and work quality, we carried out three deployments spanning two years with over ten thousand workers. We found that an interactive tutorial system is effective at teaching both novices and experts new skills. Moreover, with the right guidance, even novices were able to accomplish high-quality edits.

We offer three contributions to human-computer interaction research: First, we propose a crowdsourcing platform design that combines learning and labor to benefit both the workers and the requesters. Second, our studies extend existing research on motivating crowdworkers finding that real-world context drives workers to more closely follow directions. Finally, we offer guidelines for how platform designers can combine learning and crowdsourcing. Using these guidelines, we believe that a broad range of training environments could be transformed into crowdsourcing platforms.

RELATED WORK

In our work, we build on previous techniques used in crowdsourcing creative work [4, 8, 12, 13, 14, 16, 20] and suggest that a learning framework may further enhance creative crowd work. When the task to be crowdsourced is presented in the context of learning new skills, workers should be less likely to cheat or produce poor quality results; doing so not only harms the requester, it also subverts the learning objectives of the worker.

Learning is now increasingly happening online and opportunistically. To address this growing need for in-context help, some researchers have built interactive tutorial systems that lead users through tasks one step at a time [2, 17], while others have used game mechanics to incentivize users to keep learning [15, 7]. Games are not only useful for learning but also for crowdsourcing. There is an entire branch of the gaming community working on "serious games" or "games with a purpose." These are games that have another purpose beyond entertainment. For example, the ESP game was used to tag hundreds of thousands of images to improve image search [19], while FoldIt was used to discover new protein structures, which can be useful in creating better medication [6]. Like traditional crowdsourcing platforms, many games with a purpose ask users to do tasks that are easy for humans and difficult for computers. But beyond learning how to play the game and win, players do not learn skills that they are likely to use outside of the game. Consequently, while workers may enjoy the game experience, they are left with few marketable skills. For instance, FoldIt educates its players about the basics of amino acids and protein folding, but the target task does little to broaden skills that might be of use in biochemistry professions.

While there have been many different approaches to online learning and many different approaches to crowdsourcing complex tasks, to our knowledge, Duolingo[5] is the only project aside from ours that has combined the two and embedded crowdsourcing inside an environment for learning foreign languages. To date, most of Duolingo is devoted to traditional language instruction and the crowdsourcing task is presented as an optional translation feature, where users are offered a chance to test their language skills by helping translate documents from the Web. Since the project is yet to publish results, we don't know how well this approach works for both mastering a foreign language or for creating usable Web translations. In this paper we show evidence that a similar approach works well for learning how to edit images and for improving images supplied by real-world organizations.

SYSTEM OVERVIEW

In three deployments spanning two years, we explored how to combine crowdsourcing and learning. We built a novel platform, LevelUp for Photoshop, to support our exploration. It includes two parts (see Figure 2). First, users learn skills through interactive tutorials presented as "missions" and organized into levels. Second, at the end of each level, users test their skills in challenge rounds filled with images supplied by requester organizations.

Part I: Interactive tutorials

To make the learning experience seamless, we built LevelUp for Photoshop as a panel inside of Adobe Photoshop. All tutorials were created based on lessons from existing instructional books and were ordered by increasing difficulty level. Each tutorial contains a series of steps that guides players through common image improvements, such as removing red eye, straightening, or cutting out objects. To turn a static step-by-step tutorial into an interactive one, we associate each step with a user action, such as selecting a tool, creating a mask, or

[5] duolingo.com

LevelUp for Photoshop Platform

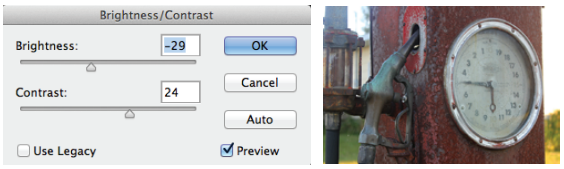
Part I: User learns skills with *interactive tutorials*

Level 1 - Mission 2: Brightness and Contrast

Step 1: Open a dull looking photo. Try our sample image to start.

User: 

Step 2: Open the Image> Adjustments>Brightness/Contrast dialog. Click on the "Auto" button to automatically adjust the image. Play with the sliders to adjust brightness and contrast separately.

User: 

Mission Completed! You have earned 30 points!

Figure 2. LevelUp for Photoshop combines step-by-step learning with real world practice through interactive tutorials and challenge rounds

saving the image, and when the user performs the action, the tutorial continues to the next step. At any given time, only one step is visible and the user must complete it to see the rest of the tutorial. By displaying one step at a time and progressing only when the user completes the step, the system gives users immediate feedback on their progress. Our system does not offer subjective feedback on the quality of image improvements. That remains future work.

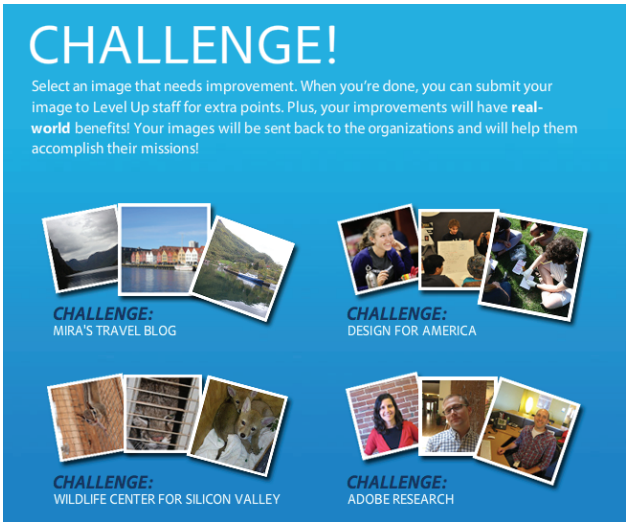
Each tutorial offers sample images, but users can also use their own images. The missions have to be completed in order, but can be repeated at any time. Each mission is associated with points, and the number of points increase with the difficulty of the mission. When they successfully complete missions, users can also earn badges, such as the "Dead-Eye Badge" for removing red eye with a single click per eye and the "Surgeon General Badge" for cutting out an object quickly. Badges such as the "Welcome Back" badge encourage players to come back a different day and continue editing images. The players can share their accomplishments through Facebook and Twitter and can compete against other players through a daily leaderboard available on the LevelUp for Photoshop website where users can also report problems and download the extension.

Part II: Challenge rounds

To support crowdsourcing we created a *challenge round* for each level, which offers images submitted by real-world organizations. We were inspired by the apprenticeship educational model, where students (e.g. aspiring chefs, hair stylists, or doctors) work on real problems as part of their training.

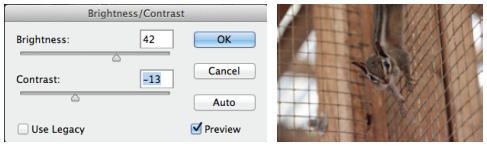
Part II: User tests skills in *challenge rounds*

Level 1 - Challenge: Images from four requesters

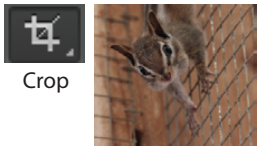


User: 

Step 1: This photo needs a lot of work. Start with the following improvements:
- Sharpen the image (25 pts)
- Adjust brightness (25 pts)

User: 

Step 2: Other ideas? Make additional improvements, and we will score the image for more points.

User: 

Step 3: Upload the image. Wildlife Center for Silicon Valley may use this image for their website, brochures, and newsletters.

Unlike the interactive tutorials, the challenge round doesn't prescribe a set of steps for improving an image. Instead, it suggests improvements. For example, the challenge round for Level 1 (Challenge 1) suggests adjusting brightness and contrast and sharpening (see Figure 2). Additionally, each challenge round also lets the user improve the image in other ways and offers extra points for additional edits. Once the user edits an image, he is given the option of uploading the image for review by LevelUp for Photoshop staff for additional points and dissemination to requesters.

Requesters

In this paper we focus on the worker experience and thus did not implement a requester interface. However, we envision an interface similar to those of existing photo editing services (Tucia,[6] Photo Editing Company[7]). These interfaces require requesters to specify how their images should be improved. The improvements specified by the requesters can map to the suggestions given to workers in the challenge rounds. Requesters with little image editing knowledge may have difficulty specifying how an image should be improved. In such situations the challenge rounds can offer suggestions that are common across many images, such as adjusting lighting, cropping, and straightening.

Implementation

The LevelUp for Photoshop panel was implemented as an extension to Adobe Photoshop in ActionScript 3.0 using Adobe Extension Builder. To install LevelUp for Photoshop, users have to use the Adobe Extension Manager, which manages installation of extensions for all Adobe products. The panel includes twelve tutorials divided in three levels. To make the tutorials interactive, LevelUp continuously monitors Photoshop events (e.g., which tool was selected, which actions were performed). If a player successfully performs the requested behavior, the panel automatically proceeds to the next task. This gives players instant feedback on their performance.

To simplify deployment we bundled all images with the extension and thus had to limit the number of images available in the challenge rounds. Each challenge round offered the user a choice of editing up to 64 images provided by four different organizations. If extended to dynamically download images after the platform is installed, the challenge rounds can support any number of organizations and images. Uploaded images were stored on Amazon S3. To support a public leaderboard, we used the Nitro game engine.[8]

DEPLOYMENT 1: LEARNING NEW SKILLS.

In our first deployment, we focused on the interactive tutorials and wanted to understand how well they taught our workers new skills. Thus, we removed the challenge rounds and deployed a version of the panel that only included tutorials. While previous work has shown that interactive tutorials are effective [2, 15], we wanted to confirm that our design could quickly lead users to learn new skills. For this deployment, the panel included tutorials for: removing red eye, improving a smile by whitening teeth, removing wrinkles, objects, and other glitches, straightening a photo, masking, adjusting lighting, replacing colors, and improving composition.

We released LevelUp for Photoshop in September 2011 and collected data for one year. To evaluate this design, we used four sources of data. First, we instrumented the game panel and measured how many missions were completed by each player. Second, we surveyed players monthly to gather their feedback. In total, we collected 470 survey responses. Third, we carried out detailed interviews with six of our players (4 women, 2 men, ages 16-68, 2 novice, 4 advanced) to learn more about their game experience. Last, we analyzed data from the Adobe Product Improvement Program, which tracks how Adobe Photoshop users use the software over time. This final data source allowed us to compare user behavior before and after the game for 218 players (63 completed the game). This dataset included people who used Adobe Photoshop for at least 7 days before installing LevelUp for Photoshop and who had used the software for at least 7 days following installation of the game. Participation in the program is voluntary, which is why we were not able to analyze this data for all players.

[6] tucia.com
[7] photoeditingcompany.com
[8] bunchball.com

Results

During the yearlong deployment, 5350 people downloaded and installed the game and completed at least one mission. Of the 5350 people, 62% completed Level 1, 49% completed Level 2, and 35% completed Level 3. Much of our survey and interview feedback was very positive. Of our 470 survey respondents, 86% reported enjoyed playing the game, 78% agreed that the game helped them learn something new, and 65% felt that they had learned the basics of photo editing. When asked why they stopped playing the game, 40% reported that they had finished the game. Only 4% reported that it was too easy or that they knew what it was trying to teach them, and 5% reported that they were just trying it out.

Was the interactive tutorial format effective?

In interviews users reported that the contextual interactive structure of the game was very useful. A lack of appropriate vocabulary can often be a stumbling block for people searching for tutorials, especially those who are just getting started with the software. Novices also reported that the step-by-step nature of the game design helped them focus on one tool at a time, rather than being overwhelmed by the entire interface. One user mentioned: "It got me to try things and gave enough instruction that I was able to rapidly make progress. Usually I get lost trying to find the item I am looking for."

Both novice and advanced users reported that they initially played the game to win points, but then revisited the game panel to refresh their memory and use the tutorials on their own images. This use of the panel was not intended but was an unexpected benefit.

Did players learn new tools and techniques?

In our surveys, players reported that some of their favorite aspects of the game included learning new tools and techniques, finding more efficient ways to perform tasks that were difficult, and the fact that it was fun to learn.

In interviews we heard similar things.

> P1: *"A lot of it is just accidental discovery. I blame myself, because I haven't been as diligent about reaching out and finding the resources. This is why I enjoyed this game. It was fun and the exercises were brief enough not to cause one to become frustrated. They introduced me to features and functions that I never knew existed."*

We also examined software use log data describing the behavior of 63 users before and after playing the entire game. We

found that everyone tried new tools they had not used before installing LevelUp for Photoshop. After finishing the game, 83% of the users continued to use the new tools they were introduced to in the game.

Were players motivated by the points and badges?
In the forums, we received many questions about how to achieve certain badges, so while it's difficult to ascertain what percentage of players were motivated by the points and badges, it seems clear that many were incentivized by these game mechanics. During the yearlong deployment there were a number of prizes that were tied to points, which offered additional motivation, but despite the fact that the prizes were limited to the US and Canada, as many as 60% of the players were from other countries.

The design in our first deployment proved to be a useful way to teach new skills, but it needed modification before we could achieve our crowdsourcing objectives.

DEPLOYMENT 2: ADDING CHALLENGE ROUNDS

In our second deployment, our goal was to test whether skills taught in a tutorial could be applied to more open ended real-world tasks using a variety images. In this study, we tested our entire platform including both the interactive tutorials and the challenge rounds. However, we did not immediately open up the platform to real-world organizations. Instead, we presented the challenge rounds as another part of the game. We were curious to see whether on their own the challenge rounds were an engaging part of the platform.

Methodology

Deployment 2 started in September of 2012 and lasted four months. We analyzed data from three sources: (1) behavioral logs, which included the number of images edited and the types of edits that were performed, (2) qualitative assessments by workers from MTurk, who compared the original images to those uploaded by users of our platform, and (3) qualitative assessments by expert raters, which examined the quality of the edited images. Additionally, all players filled out a short survey specifying their experience level, gender, and locale.

For the qualitative assessments we created a subset of images to analyze by randomly sampling 552 users and selecting all of their images (2383 in total).

We collected task independent feedback from workers on MTurk using a paired comparison interface that showed two images (the original and an improvement) and asked them to select the image they believed would be more useful for a professional website. The placement of the original was randomized. Each job included twenty image pairs and paid $0.10. Two of the twenty pairs were used as golden standards. The golden standard pairs were created by expert raters. Each pair was evaluated by five different workers.

Inspired by previous work in creativity assessment [1, 9, 11], we also assessed the photo improvements systematically using expert raters who knew a lot more about the game environment and the suggested improvements. The expert raters gave each image two ratings, one for usefulness (1-3) and one for novelty (1-3).

The usefulness scale measured whether the edited image would be more valuable to the requester than the original image 3=*more useful than the original*, 2=*as useful as the original*, and 1=*less useful than original*. The novelty scale described the complexity of edits performed by the users. Users who used tools beyond those suggested by the challenge round, such as blurring part of the image or changing it from color to black and white, were given a *high novelty* rating (3). Users who stuck to the tools suggested by the challenge round but created novel effects were given a *medium novelty* rating (2). Finally, users who followed the directions and did exactly what they were asked performing only the suggested edits for each challenge round were given a *low novelty* rating (1). Figure 3 shows several images and the corresponding usefulness and novelty ratings.

As previous scholars have acknowledged [18], there are limits to creativity assessments as expert ratings are subjective. For this reason, we relied on established means for reaching inter-rater reliability [5]. First, the two raters rated 11% of the images. Since the inter-rater reliability between the two raters was high (Cohen's Kappa=0.96 $p<0.001$ for novelty and 0.92 $p<0.001$ for usefulness), we divided the remaining corpus between the two raters. Both raters were members of the research team.

All results reported as statistically significant used a Mann-Whitney U test with Bonferroni corrections to adjust for multiple comparisons and Type I error.

Results

3512 people (35%female) from 82 countries (35% from the US) participated and submitted 8068 images. Each image in our dataset was improved by at least 8 people. The majority of users (78%) reported that they were just starting to use Photoshop (novices). 11% said they had between 1 and 5 years of experience with the software (intermediates), and 4% said they had more than 5 years of experience (experts). 7% did not give an expertise level. Using the file names, we confirmed that 98% of the submitted images were improvements of the images embedded in the challenge rounds.

Worker performance: quantity

Consistent with our observations in Deployment 1, 74% of users completed Challenge 1, 57% of users completed Challenge 2, and 39% completed Challenge 3. Surprisingly, the number of images submitted in each round did not follow the same pattern. 48% of the images were submitted in Challenge 1, 26% were submitted in Challenge 2, and 26% were submitted in Challenge 3.

Of the 2613 users users who completed Level 1, 92% submitted at least one image. 31% submitted one image, 25% submitted two images, 20% submitted three images, and 24% submitted 4 or more images. Two people improved all 192 available images. Ten people improved over 100 images each. These ten motivated people submitted 21% of all of the submitted images, which is substantially lower than findings for MTurk, which show that the motivated individuals who do all the HITS for a task end up accomplishing almost half (46%) of all the work [10].

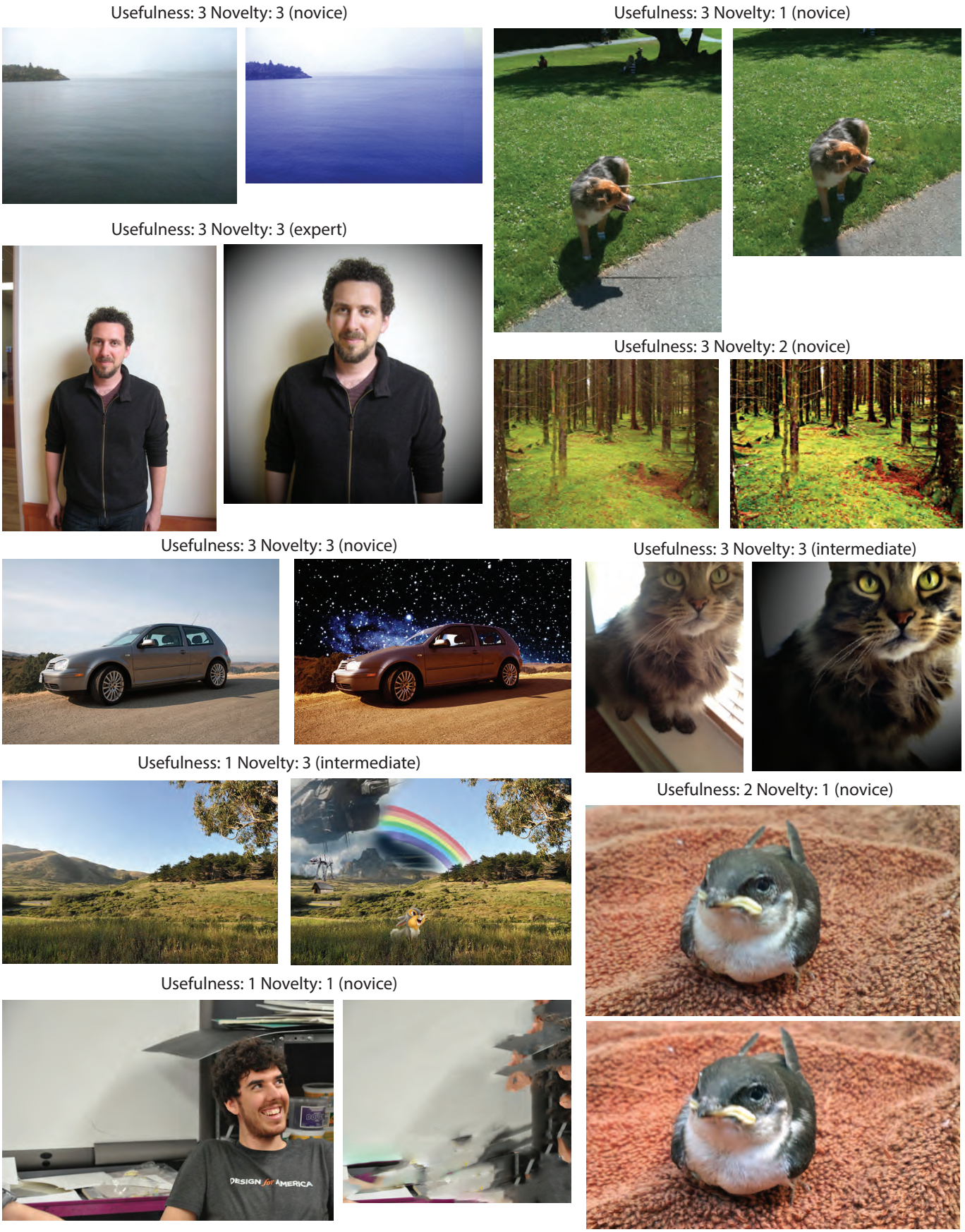


Figure 3. Here are some of the images submitted by our workers and their ratings for novelty and usefulness.

We analyzed the behavior of the 84 users who completed all challenge rounds and submitted at least ten images and found the following three patterns. 47% of users worked on roughly the same number of images in each challenge. 32% worked on a lot of images in the first challenge and almost none in the subsequent challenges. 16% worked on almost no images in the first two rounds but a lot in the final round. We believe that those who only worked on images in the first challenge were more interested in the interactive tutorials and seeing what the game had to offer. Those that only submitted images in the last challenge round may not have felt ready to tackle the challenge round until they had finished all of the tutorials. Finally, 39% had a strong preference among image categories, primarily working on images in one or two categories.

There was a strong preference for editing nature images over all other types of images. 30% of the edited images were outdoor nature scenes. 22% were animal images, 23% were people images and 24% were car images. We found no correlation between the expertise, gender, or age workers and the number or types of images improved.

Worker performance: quality

We used two methods to assess quality: (1) task-independent assessments from workers on MTurk (turkers) and (2) expert evaluation of novelty and usefulness.

Assessment by turkers

We collected 14760 total responses from 162 turkers. We excluded responses from workers who failed the golden standard comparisons more than 25%, which resulted in 6360 responses from 82 turkers for 2636 images. We found that for 25% of the submissions all turkers agreed that the submitted image was better than the original. For 37% for the submissions the majority of the turkers agreed that the submitted image was better than the original. The results were similar across player expertise. The turkers preferred improvements performed as part of Challenge 1 and Challenge 2. The majority of turkers agreed that 41% of submissions in Challenge 1 and 40% of the submissions of Challenge 2 were better than the originals, while in Challenge 3 only 25% were deemed better than the original by the majority of turkers. This may be due to the suggested edits. In Challenge 3, the suggestions were to convert the image to a square and remove glitches. A square composition is less common, and thus, the turkers may have preferred it less.

Assessment by expert raters

Of the 2383 images that were rated for usefulness and novelty, 69% were rated as more useful than the original, 15% were rated as less useful and 16% were rated to be about the same. The expert ratings are higher than those by the turkers, because the experts gave a high rating if a worker completed the task he was asked to do, even if the edit was not significantly better than the original. For example, if a worker cropped an image as a square in Challenge 3 and did not do anything too objectionable, such as crop across a face, the submission was given a rating of 3 for usefulness, even if the composition was not significantly improved by a square crop. 44% of users did not submit any images that were rated less useful than the originals. 20% of users did not submit even one image that was rated as useful. Previous work [4, 10] reported similar results with 10-30% error rate on Amazon Mechanical Turk for HITs without quality control measures. Figure 3 shows some of the more exceptional work submitted by the workers. Many users tried tools and techniques beyond those suggested by the challenge rounds. There were many who applied painterly effects to images, blurred the background or replaced some portion of the image. Some workers interpreted our instructions to do additional edits more liberally and attempted to do whimsical and funny edits. For example, some inserted cartoon characters, while others tried to communicate by adding text to the images.

The role of experience

When we compared users with different experience levels, we found that expert users received significantly more highly useful ratings (μ=2.833) as compared to novice (μ=2.54) and intermediate users(μ=2.43) ($p<0.0001$ for both). 86% of images edited by experts were rated as more useful than originals (69% for novices, and 62% for intermediate) and only 3% were rated as less useful (15% for novices and 19% for intermediates). This suggests that our rating rubric is consistent with real-world expectations. Similarly, experts received significantly more high novelty ratings (μ=1.55) as compared to novice (μ=1.21) and intermediate users (μ=1.27) ($p<0.0001$ for both). This is not surprising because experts should know how to use many more tools than novices and thus can improve images with a larger variety of techniques. 27% of the images edited by experts received a high novelty rating because they used tools beyond those suggested by the challenge rounds (9% for novices and 11% for intermediates).

Since our worker population is largely made up of novices (78%) and our training platform holds the most potential for those just getting started, we also analyzed the performance of novice users. Of the 552 users whose images we rated, 436 (79%) were novices. Of the novices, 339 (78%) submitted at least one image that was rated as useful. 184 (42%) only produced images that were rated as useful. These results are very encouraging, because they show that an interactive tutorial environment paired with less structured challenge rounds provides enough scaffolding to produce high quality work even from users who have little prior experience.

Surprisingly, in the set of images that were rated, we were not able to identify *cheaters*, i.e. people who were just trying to get points. We defined cheaters as people who got more than 50% of their images rated with 1,1 (i.e. less useful than the original and low novelty).

Challenge 1 vs Challenge 3

We found that significantly more images received "more useful" ratings in Challenge 3 than in Challenge 1 (76% in Challenge 3 vs 66% in Challenge 1 $p<0.0001$). We believe this is the case because only 40% of the users got to Challenge 3, and presumably these users were more motivated to learn and do a good job. In Challenge 1 many people might have been trying out the game and exploring its features, without trying to do a good job.

Wildlife Center for Silicon Valley (WCSV) provides high quality care and rehabilitation of injured, sick, and orphaned wildlife within the Silicon Valley Community. Through the dedication of approximately 120 volunteers, they care for over 4000 birds and mammals from over 150 species each year. "High quality images are critical for helping us obtain funding and recruit volunteers." - Janet Alexander, Director of Operations The images you edit will be sent to WCSV for their website, brochures, and newsletters.
Design for America (DFA) is a national training program for students interested in creating solutions to local challenges in their community in health, education, and the environment. With the help of more than 1200 students and professionals, they have touched the lives of 1000s of community members. "High quality images are critical for helping us recruit student and professional volunteers and raise money to support our programs." - Elizabeth Gerber, Founder The images you edit will be sent to DFA for their website, brochures, and newsletters.
Mira, LevelUp for Photoshop staff and player, wants to start a travel blog for family and friends. She just returned from Norway and has a large collection of images that she wants to put in her blog. "I try to capture the feeling of being there, but my photos often turn out dull. I know more people will read and enjoy my blog if I show nice images." - Mira The images you edit will be sent to Mira to help her start her blog.
Adobe Research is collecting a dataset of portrait photos for use in new research projects. "We want to build software that can fully interpret portraits of people and thereby enable creative manipulation of portraits that is both simple and fun. To do this, we need a large dataset of portrait photos in all their variety." -Jon Brandt, Principal Scientist The images you edit will be sent to Adobe Research for use in projects.

Table 1. The text describing the requester organizations.

We also found that images from Challenge 3 received significantly fewer high novelty ratings (10% vs 15% for challenge 1 $p<0.0001$). This is somewhat surprising, because workers could have used tools they learned in earlier levels to make additional edits beyond those suggested in Challenge 3. However, fatigue effects can affect creativity, so if users were more tired towards the end of the game, they may have been less willing to do additional creative edits.

While we observed a preference for editing nature images, we did not find the usefulness or novelty ratings for nature images to be significantly higher.

DEPLOYMENT 3: EXPLORING MOTIVATION STRATEGIES

In our third deployment, we added real-world context to the challenge rounds by embedding images from requesters. We curious whether certain types of organizations (e.g. non-profit vs for-profit) would receive better treatment and whether through the design of the game we could influence users to edit more images and do a better job. We collected images from four different organizations: a non-profit wildlife center, a non-profit student organization, a travel blogger, and a for-profit research organization that was collecting images for research projects. We placed these images in the challenge rounds and deployed two different versions of the game to test our two different designs.

Methodology

Deployment 3 started in March 2013 and lasted six months. In the *purpose* design, we included a lot of information about each organization, including a quote describing how the images will be used by the organization. In the *control* design we only listed the name of the organization. Table 1 shows all of the text we used in the purpose condition.

We randomly sampled 48 images from the images provided by each organization and placed them in the three challenge rounds. Unlike in Deployment 2, where each image category was pretty consistent, in this study the images were more varied, because the organizations provided the images they wanted to use in their marketing materials. So for example, the images from the wildlife center included images of wildlife as well as images of volunteers working in the wildlife center.

We analyzed data from five sources: (1) behavioral logs, which included the number of images edited and the types of edits that were performed, (2) qualitative assessments from workers on MTurk, (3) qualitative assessments from experts, (4) qualitative assessments from two requesters on how the edited images compare to the originals, and (5) surveys assessing the user experience.

To get feedback from the requesters, we used the same paired comparison interface that we showed to the workers on MTurk. We asked requesters to select the image they believed would be more useful to their organization. As with Deployment 2, we used the Mann-Whitney U test for all reported p values and applied Bonferroni corrections to address multiple comparisons.

Results

1290 people (648 purpose condition, 642 control condition, 27% female, 57% male) participated in Deployment 3 and worked on 2606 images. In the purpose condition we had 69% novices, 12% intermediates, 8% experts, and 11% who did not specify expertise. In the control condition we had 65% novices, 14% intermediates, 9% experts, and 12% who did not specify expertise.

We found no difference in the total number of images submitted between conditions. Similar to what we observed in Deployment 2, users preferred editing outdoor landscape images over other types of images. While we expected non-profit organizations to receive more attention from workers, we found little evidence to support this hypothesis. In the surveys users reported that they chose images not based on organization, but based on relevance to their lives. So, people edited outdoor landscape scenes more often, because many of their personal images were of outdoor scenes.

We found no significant difference between conditions in the qualitative turker assessment. Similarly, we found no significant difference between conditions or between organizations

in the expert ratings for usefulness. But we did find a significant difference in the novelty ratings between the purpose and control conditions ($p<0.0001$). The control condition had a higher average novelty rating (control: μ=1.37) than the purpose condition(μ=1.22). In the purpose condition 87% of images received a low novelty rating, while in the control condition 80% of images received a low novelty rating. We believe that people under the purpose condition were less willing to do more than the suggested edits because they wanted to do what was asked by the organization.

Requester feedback to the submitted images was positive. The Wildlife Center for Silicon Valley rated 76 submitted images. They rated 60% of the submitted images as better than the originals. Of the images that were submitted by novices (50 (66%)), they rated 92% as better than the originals.

Design for America rated 470 submitted images. They rated 20% of the submitted images as better than the originals. Of the images that were submitted by novices (271 (58%)), they rated 34% as better than the originals.

We collected 35 surveys and received qualitative feedback on the design of LevelUp for Photoshop and what could be improved. When asked what they liked most about the challenge rounds, workers listed new ways to learn ("I got to apply what I had just learned," "It's a nice ramp up from the previous lessons,""I like that I get to test the skills that I've learned.") and real-world impact ("It offers a sense of achievement beyond badges and points," "Having the opportunity to help real people/organizations out," "Working on real problems.") When asked what they liked least, workers reported that some of the images had poor resolution, the task felt repetitive, and that there was no feedback from the organizations. When asked whether they preferred the missions or the challenge rounds, almost all users (except for one) said that they preferred the missions or had no preference between the missions and the challenge round.

DISCUSSION

Our studies show that interactive learning can be combined with crowdsourcing to enhance both learning and worker performance. We found that an interactive tutorial system is effective at teaching both novices and experts new skills. Moreover, with the right guidance, even novices were able to accomplish high-quality edits. Furthermore, offering real-world context about a requesting organization motivates workers to more closely follow directions.

Design Implications for Creative Crowdwork

Designers of crowdwork platforms face many challenges in designing interactions that positively influence the requesters' and the workers' experiences. Our study highlights several design implications for crowdsourcing complex creative work. First, and perhaps most importantly, learning platforms can be an effective way to produce high quality creative work. If a particular creative skill is hard to find in traditional marketplaces, one solution is to offer workers a chance to learn this skill in exchange for producing work. This design approach can benefit both requesters and crowdworkers.

We also illustrate how simple it can be to adapt online learning tutorials to work with crowdsourcing systems. Many online learning platforms offer students a way to practice their skills, often through sample use cases that serve no purpose beyond pedagogy. In most photo editing tutorials, for example, students are given stock photos to improve. In our system we illustrate how stock photos can be replaced with third party images that are in genuine need of image enhancement. Similar substitutions could be made for other learning systems. We argue that this paradigm not only helps produce good crowdsourced work, but it also enhances the user experience and motivations of the workers.

Further, our challenge round design illustrates ways to capture labor from both novice and expert workers. For novices, it is important to provide specific, structured objectives. For experts, it is important to also make tasks somewhat open-ended, so they can utilize additional skills as appropriate. In our challenge rounds, experts were given the opportunity to make additional improvements to the image not specified by the requester. While there is a chance some workers might take this opportunity to overreach and perhaps make changes that are not appropriate, our experiments show that these behaviors are reduced when the needs of the requester are made explicit. Indeed, our findings also add further support to the notion that a sense of purpose can enhance motivation. When workers were given descriptions of the requester, they followed the instructions more closely and made more appropriate adjustments.

Future work

In future work we plan to extend the challenge round design to incorporate subjective feedback. Offering feedback is an integral part of learning and has been shown to be effective in crowdsourcing [20]. One possible design is to ask workers to rate images that others have submitted after they submit their own edits. Such an approach scales well, and potentially has both pedagogical benefits (reviewing others' work is an important part of instruction) and motivational effects (individuals are motivated by both opportunities to collaborate and to compete).

Limitations

Our worker pool was limited to those who voluntarily chose to use LevelUp for Photoshop, and we benefitted from tying our platform to a popular application. Not all tasks are amenable to being embedded into an existing application, and requesters may be limited to working with platforms such as MTurk. We believe that our approach could be applied on top of existing platforms, by integrating learning into tasks, but more research is necessary to further understand how payment would interact with learning motivations.

Many in the crowdsourcing community have focused on improving cost and throughput [3], but evaluating the cost and throughput of our platform was beyond the scope of this paper. Throughput for an in-app voluntary platform like LevelUp for Photoshop is heavily affected by marketing efforts and product release schedules. As a point of comparison, to

improve all of the images edited in our experiments would have cost over $3200 on Tucia.

CONCLUSION

We present a novel platform that combines learning and crowdsourcing to benefit both workers and requesters. To understand the factors that influence learning outcomes and work quality, we carried out three studies over two years with over ten thousand workers. We found that:

- Workers enjoyed the interactive tutorial environment and learned new photo editing skills. All workers who completed all of the tutorials learned at least one new tool that they had not used prior to installing LevelUp for Photoshop, and 83% continued to use the new skills outside of the interactive tutorial game
- Even novice Adobe Photoshop users were able to complete basic photo editing tasks and improve images. 69% of images edited by novices were considered more useful than the originals as determined by expert raters blind to the experimental conditions. Additionally, the requesters who provided the images rated up to 92% of the images submitted by novices as more useful than the originals.
- Workers preferred work that helped them practice the skills they found most useful. For example, workers who wanted to improve their own landscape photos chose to work on crowdsourced tasks that involved outdoor scenes.
- Highlighting the purpose of the crowdsourced work changed the behavior of the workers. When the requester's origin was known and the purpose of the image enhancements was made salient, workers followed instructions more closely and made more appropriate adjustments.

ACKNOWLEDGMENTS

We thank all of the LevelUp for Photoshop players and Amazon Mechanical Turk workers. Also, we thank the Wildlife Center for Silicon Valley and Design for America for contributing images and working with us.

REFERENCES

1. Amabile, T. M. *Creativity and innovation in organizations*. Harvard Business School, 1996.
2. Bergman, L., Castelli, V., Lau, T., and Oblinger, D. Docwizards: a system for authoring follow-me documentation wizards. In *Proceedings of UIST* (2005), 191–200.
3. Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of UIST* (2011), 33–42.
4. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proceedings of UIST* (2010), 313–322.
5. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1 (1960), 37–46.
6. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., and players, F. Predicting protein structures with a multiplayer online game. *Nature 446* (2010), 756–760.
7. Dong, T., Dontcheva, M., Joseph, D., Karahalios, K., Newman, M., and Ackerman, M. Discovery-based games for learning software. In *Proceedings of SIGCHI* (2012), 2083–2086.
8. Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. Shepherding the crowd yields better work. In *Proceedings of CSCW* (2012), 1013–1022.
9. Guilford, J. P. *The nature of human intelligence*. McGraw-Hill, 1967.
10. Heer, J., and Bostock, M. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of SIGCHI* (2010), 203–212.
11. Kaufman, J. C., Plucker, J. A., and Baer, J. *Essentials of creativity assessment*, vol. 53. Wiley, 2008.
12. Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E. Crowdforge: crowdsourcing complex work. In *Proceedings of UIST* (2011), 43–52.
13. Kulkarni, A., Can, M., and Hartmann, B. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the CSCW* (2012), 1003–1012.
14. Lewis, S., Dontcheva, M., and Gerber, E. Affective computational priming and creativity. In *Proceedings of SIGCHI* (2011), 735–744.
15. Li, W., Grossman, T., and Fitzmaurice, G. Gamicad: A gamified tutorial system for first time autocad users. In *Proceedings of UIST* (2012), 103–112.
16. Morris, R. R., Dontcheva, M., Finkelstein, A., and Gerber, E. Affect and creative performance on crowdsourcing platforms. In *Proceedings of ACII* (2013).
17. Pongnumkul, S., Dontcheva, M., Li, W., Wang, J., Bourdev, L., Avidan, S., and Cohen, M. F. Pause-and-play: automatically linking screencast video tutorials with applications. In *Proceedings of UIST* (2011), 135–144.
18. Puccio, G., and Murdock, Mary, E. *Creativity Assessment: Readings and Resources*. Creative Education Foundation Press, Buffalo, NY, USA, 1999.
19. von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proceedings of the SIGCHI* (2004), 319–326.
20. Zhang, H., Dow, S., Kraut, R., and Kittur, A. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of CSCW* (2014).