The Wikipedia Adventure: Field Evaluation of an Interactive Tutorial for New Users

Sneha Narayan
Northwestern University
Evanston, IL
snehanarayan@u.northwestern.edu

Jake Orlowitz
Wikimedia Foundation
Bryn Mawr, PA
jorlowitz@wikimedia.org

Jonathan Morgan
Wikimedia Foundation
San Francisco, CA
jmorgan@wikimedia.org

Benjamin Mako Hill
University of Washington
Seattle, WA
makohill@uw.edu

Aaron Shaw
Northwestern University
Evanston, IL
aaronshaw@northwestern.edu

ABSTRACT

Integrating new users into a community with complex norms presents a challenge for peer production projects like Wikipedia. We present The Wikipedia Adventure (TWA): an interactive tutorial that offers a structured and gamified introduction to Wikipedia. In addition to describing the design of the system, we present two empirical evaluations. First, we report on a survey of users, who responded very positively to the tutorial. Second, we report results from a large-scale invitation-based field experiment that tests whether using TWA increased newcomers' subsequent contributions to Wikipedia. We find no effect of either using the tutorial or of being invited to do so over a period of 180 days. We conclude that TWA produces a positive socialization experience for those who choose to use it, but that it does not alter patterns of newcomer activity. We reflect on the implications of these mixed results for the evaluation of similar social computing systems.

Author Keywords
newcomer socialization; gamification; systems design; systems evaluation; peer production; Wikipedia; online communities.

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation (e.g., HCI): Group and Organization Interfaces – Computer-supported cooperative work

INTRODUCTION

Social computing systems and peer production communities that aggregate voluntary contributions depend critically on recruiting and retaining new users [30]. Since no user will contribute indefinitely, online collaborative projects must successfully mobilize newcomers in order to maintain their community. However, in order to successfully make high-quality contributions and avoid censure, new users must quickly learn community norms. As inexperienced users inevitably violate norms, the impetus to recruit newcomers can be in tension with the desire to maintain quality.

An interstellar journey through Wikipedia. Full of real skills, tips, tools, tricks, helpers, sources, rewards, and support.

Figure 1. Welcome page shown to users who arrive at The Wikipedia Adventure for the first time.

Wikipedia provides a well-known case of this dilemma in social computing. Since a short contribution history is an excellent predictor of vandalism in Wikipedia [38], established community members often delete or "revert" newcomer contributions. This demoralizing experience drives many good-faith newcomers away [22, 25, 33] and has contributed to an overall decline in Wikipedia's active editors since 2007 [22, 44]. Making matters more difficult for newcomers, Wikipedia's norms, procedures, conventions, and policies have expanded considerably since the inception of the community [9]. While a growing body of design research aims to overcome these challenges [23, 33], existing systems frequently rely on the helpfulness of veteran editors (a limited resource), and significant initiative from newcomers themselves.

We present a novel system called The Wikipedia Adventure (TWA): an interactive tutorial that provides an introduction to editing Wikipedia. Unlike most prior systems designed

CSCW '17, February 25 - March 01, 2017, Portland, OR, USA
ACM 978-1-4503-4335-0/17/03.
http://dx.doi.org/10.1145/2998181.2998307

to socialize new users on Wikipedia, TWA creates a structured, interactive experience that guides newcomers through critical pieces of Wikipedia knowledge: editing using wikimarkup (the code that editors use to format text), communicating with other editors, and learning basic community policies. It also incorporates elements of gamification in an attempt to increase the motivation, engagement, and enjoyment of newcomers as they learn about the community.

After describing TWA, we evaluate it in two ways. First, we report a survey that assesses how new users perceive the system's design and tone. We then conduct a randomized controlled field experiment in which we invite new Wikipedia editors to use the system and measure its effect on their subsequent contributions using multiple parametric and nonparametric techniques. One of these techniques is two-stage least squares regression which we use to estimate the effect of playing TWA (conditional on having been invited to do so) on the number of future contributions.

We find that survey respondents perceived TWA as engaging and well-designed for newcomers. At the same time, our field experiment shows that deploying TWA does not alter newcomer contribution patterns. These results imply that the design principles of TWA are sound but that the system does not produce the expected impact. We propose multiple potential explanations for this null effect and suggest that they may point to a gap in existing literature on newcomer socialization. In addition to the empirical findings, the study also contributes one of the first randomized controlled studies of the effect of a gamified orientation system as well as the first application of two-stage least squares analysis in social computing research.

BACKGROUND

Newcomer Socialization in Online Communities

The norms and routines of organizations and communities often appear opaque to newcomers. Failing to communicate the skills that a newcomer needs in order to effectively contribute can lead to frustration and alienation for newcomers, while also breeding distrust of newcomers among established members. For online communities that depend on voluntary contributions from their users, socializing newcomers is among the most crucial tasks [30].

Prior research on newcomer socialization in organizations distinguishes between *individualized* and *institutionalized* socialization tactics [30, 45]. Individualized socialization is informal, akin to on-the-job learning, and is directed by newcomers themselves. Institutionalized socialization is more collective and formal, with the aim of providing a uniform set of experiences. In conventional firms, there is a broad consensus that institutionalized forms of socialization are more effective in retaining new members [5]. Institutionalized socialization techniques facilitate newcomers joining organizations by increasing self-efficacy, providing role clarity, and instilling a sense of social acceptance which, in turn, leads to better performance and higher commitment [5].

While several online communities and social computing systems use institutionalized socialization, large peer production communities, including Wikipedia, have relied almost exclusively on individualized socialization techniques, and typically require users to figure out what they need to know in order to contribute to a project [15, 17, 34]. More general, formal, and institutionalized systems for newcomer socialization can complement these existing approaches.

Gamified Onboarding

In considering strategies for effectively onboarding newcomers in online communities, we draw on recent research showing that gamification can support engagement in interactive systems. Gamification has become an increasingly popular approach within interactive system design for improving engagement in learning activities. In a meta-analysis of gamification studies in computer science, HCI, and eLearning, Hamari et al. [26] identified 10 common motivational affordances of gamified systems: points, leaderboards, achievements/badges, levels, stories/themes, clear goals, feedback, rewards, progress, and challenges. They showed that in a majority of cases they reviewed, these features led to positive learning outcomes and enhanced enjoyment among participants.

However, gamification has limits. Work in psychology [12] and behavioral economics [18], as well as studies of gamified systems [27], have highlighted the potential demotivating effect of gamification. In particular, competition-based incentives which are central to gamification affordances like leaderboards have been shown to undermine participants' motivation.

The effects of gamification in non-gamified contexts over time remain largely unstudied. As a result, the question remains open as to whether a gamified tutorial that effectively introduces a particular task will have any effect on participants' motivation to perform that task outside of the tutorial. On one hand, a gamified introduction to a task might increase positive affect, confidence, and self-efficacy which might in turn increase subsequent participation. On the other hand, the social psychology literature on crowding out [12], has shown that shifting incentives, particularly the removal of extrinsic incentives, can decrease levels of motivation overall.

The Challenges of Becoming a Wikipedian

A mature community like Wikipedia poses particular challenges for onboarding newcomers. Wikipedia has seen a massive decrease in newcomer retention over the past decade. Between 2006 and 2010, Wikipedia's retention rate of newcomers acting in good faith (i.e. those whose initial edits showed a desire to contribute productively) dropped from 25% to 5% [22]. Research has suggested that this drop in retention is due in part to higher rates of negative socialization experiences like receiving warning messages or having an edit reverted [25], the lack of effective socialization in the presence of increasingly formal policies and rules, and an increase in the use of automated quality control tools to enforce rules and sanction new users [9, 22].

Most new editors begin editing without any structured external guidance and, perhaps as a result, quickly adopt behaviors and roles that tend to persist over their Wikipedia careers.

Panciera et al. [37] have shown that initial rates of editing are among the strongest predictors of long-term contribution rates. That said, subsequent work has shown that contributions from extremely active users in social computing systems also change dynamically over time in ways that can be influenced by both users' experiences and technological interventions [28]. Panciera et al. and others have suggested that the creation of systems to support newcomers immediately after joining Wikipedia is important for building long-term commitment [30, 37].

Historically, Wikipedia has relied on user-initiated, individualized forms of newcomer socialization. New editors on Wikipedia frequently learn what to do by requesting help and feedback [36] and by consulting Wikipedia's help and policy pages. They also learn what not to do through warning and advisory messages they receive from experienced editors when they violate a policy or make a mistake [19].

While new Wikipedia editors select their own tasks, the increasing scope of the community policies, distrust from experienced editors, and requirements of specialized knowledge has made it difficult for them to act independently [8, 22]. In a 2010 survey, over 40% of new editors who decided to leave Wikipedia cited a lack of support or an unpleasant social atmosphere.[1] In particular, women reported that they found that contributing to Wikipedia involved a high level of conflict and that they lacked confidence in their expertise [11].

Why Gamify Becoming a Wikipedian?

A gamified tutorial has the potential to increase newcomer participation on Wikipedia for several reasons. At a minimum, tutorials seem to have helped existing active Wikipedians. In a survey of readers and editors of the French Wikipedia, regular contributors reported using tutorials when they began editing at a greater rate than occasional contributors [14]. An interactive tutorial that addresses the most critical technical and organizational topics, made available to a large number of new editors shortly after they join, might help more newcomers overcome these challenges. Taking a structured, step-by-step approach with milestones and incremental feedback can increase their sense of self-efficacy [5].

Research on experienced Wikipedians also shows that many active contributors experience several elements of gamified systems through their participation in the community. For example, WikiProjects often set project level "challenges" that encourage editors to complete a certain number of tasks in a short period of time [48]. Editors who achieve high-visibility goals in the community are acknowledged for their efforts with badge-like social awards which confer external recognition of their achievements [31]. Making such gamified elements more transparent to new editors could increase their enjoyment and lead to increased contributions.

Although prior attempts at socialization within Wikipedia have incorporated institutionalized elements, none have explicitly combined these with gamified design. The Wikimedia Foundation previously deployed a tutorial called GettingStarted which introduced participants to basic editing concepts using an interactive interface without game-like features. It showed no effect on retention and was subsequently deactivated.[2] New editors may also receive a generic welcome message from members of the Wikipedia Welcoming Committee soon after they register.[3] Typical welcome messages explain community values and provide a list of policies and resources. Wikipedia's policy pages are, in this sense, also an effort to provide institutionalized guidance about norms and routines. However, without a more structured introduction to facilitate learning about these policies, many newcomers will never read them.

Although several existing systems provide explicit and personalized direction to Wikipedia newcomers, most require extensive one-on-one interaction and effort from experienced editors. The Wikipedia Teahouse Q&A forum [33] provides a safe space for newcomers to get personalized help from experienced volunteer "hosts" and fellow newbies in a many-to-many setting. The Adopt-a-User program provides opportunities for newcomers to enroll in extended, one-on-one mentoring relationships with experienced editors [36]. The MoodBar allowed new editors to provide instant post-edit feedback that was piped to a feed monitored by experienced editors who were encouraged to step in and provide rapid assistance [10]. These examples demonstrate tensions between scale, information density, and personalization which many efforts at newcomer socialization must confront. The differences between them also suggest how a gamified tutorial might provide a scalable approach that does not feel as impersonal as a message from a bot account or as overwhelming as navigating a thicket of complicated policy pages.

While personalization is a major advantage of socialization efforts such as The Teahouse and Adopt-a-User, these systems are limited by volunteer time of hosts and mentors and the burden is still on the new user to initiate conversations and ask questions. A scalable, institutionalized effort at orientation initiated by the Wikipedia community has the potential to reach a greater number of new users more quickly and provide a framework for understanding what it means to contribute to Wikipedia.

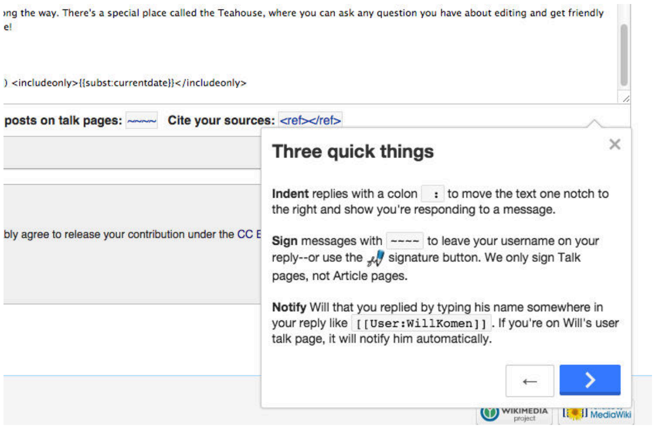
Motivated by prior research on newcomer socialization in online communities and gamification in interactive environments, we designed a system to bridge gaps in the current newcomer onboarding experience in Wikipedia. Our main research questions were:

1. *Would a gamified tutorial produce a positive and engaging experience for new Wikipedians?*
2. *Would playing the tutorial cause newcomers to contribute more?*

[1] Based on the Wikimedia Foundation's survey of former contributors, available at: https://strategy.wikimedia.org/wiki/Former_Contributors_Survey_Results

[2] See: https://meta.wikimedia.org/wiki/Research:Onboarding_new_Wikipedians/OB6

[3] An example of a welcome message can be viewed at: https://en.wikipedia.org/w/index.php?title=User_talk:Krishna_7murari&oldid=643725558



ɔng the way. There's a special place called the Teahouse, where you can ask any question you have about editing and get friendly e!

) <includeonly>{{subst:currentdate}}</includeonly>

posts on talk pages: `~~~~` Cite your sources: <ref></ref>

bly agree to release your contribution under the CC E

Three quick things

Indent replies with a colon `:` to move the text one notch to the right and show you're responding to a message.

Sign messages with `----` to leave your username on your reply--or use the signature button. We only sign Talk pages, not Article pages.

Notify Will that you replied by typing his name somewhere in your reply like `[[User:WillKomen]]`. If you're on Will's user talk page, it will notify him automatically.

Figure 2. Learning syntax through The Wikipedia Adventure.

SYSTEM DESIGN: THE WIKIPEDIA ADVENTURE

In order to address the challenges of newcomer socialization on Wikipedia, we designed The Wikipedia Adventure, a system that introduces newcomers to the nuts and bolts of editing the encyclopedia. TWA provides a positive and gamified introduction to Wikipedia by taking a new user on a guided journey through the basics of editing, communication, and community norms in order to help them develop the skills to make effective contributions.[4] While playing TWA, the user is asked to perform a series of tasks for which they are provided detailed instructions. The tasks are couched in realistic Wikipedia editing scenarios. The user's input is evaluated and they receive a prompt informing them whether they completed the task correctly. If their response was incorrect, they receive additional instruction and cues.

TWA incorporates institutionalized socialization techniques by providing a standardized, sequential introduction to the norms and policies of Wikipedia. While interactive, it does not depend on the availability, helpfulness, or intervention of existing Wikipedia editors and can therefore scale to support an arbitrary number of newcomers. The Wikipedia Adventure teaches users how to edit using wikimarkup code by using a series of pop-up boxes that point out where to click and what syntax to use in each context. Figure 2 shows how the user is given a lesson on how to edit their talk page through a pop-up box that appears next to their talk page editor.

In order to create a safe space for the user to try new things without fear of scrutiny or reprisal, TWA provides a training experience in a section of Wikipedia that is separate from existing articles. That said, this design decision also increases the risk that players will consider the edits they make through the Wikipedia Adventure to be inauthentic. We considered an alternative approach in which newcomers edited actual articles on Wikipedia. Support for this approach can be found in the theory of legitimate peripheral participation (LPP) which emphasizes that learners should be able to perceive that their initial contributions, however small, are valuable to their new community of practice [32].

Unfortunately, Wikipedia presents challenges to an LPP-based approach. Research has shown that inexperienced Wikipedia editors who edit "public" articles are likely to make mistakes that elicit powerfully demotivating reactions [25]. As a way of reducing this risk while still creating an effective system for socialization, we drew inspiration from researchers working to reconcile LPP with traditional instructional methods who faced similar challenges.

Drawing on previous work around the concept of authenticity in education [29, 43], Guzdial and Tew [21] argue that learners may derive similar benefits from performing inauthentic tasks as long as they perceive an alignment between the tasks they are assigned and the work of a community of practice they value. Moreover, they suggest that educators can facilitate such alignment through storytelling, by (for example) creating a fictional narrative context in which students can perceive their learning tasks as legitimate peripheral participation within an imagined community of practice. TWA was designed to encourage alignment in both of these ways through a number of game-like elements. While we drew significantly from the gamification literature in making design choices, we eschewed features like leaderboards to avoid potential demotivating effects associated with competition. We describe three of the gamified elements below and show how they fit with the goals of the project and the context of new users' experience of Wikipedia.

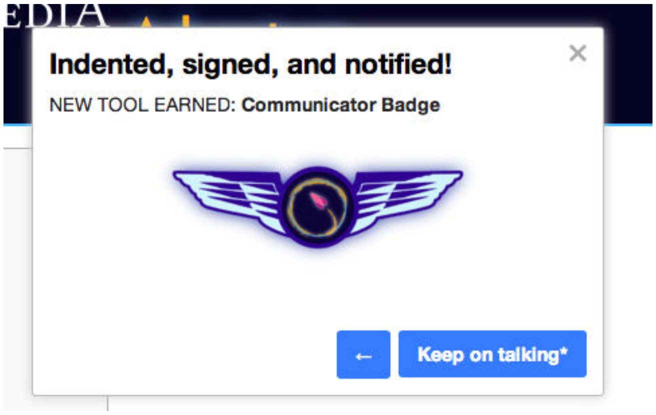
Missions

TWA is split into seven *missions* which accomplish different learning objectives tied to the five pillars of Wikipedia[5] and reflect key community rules and norms [40]. Missions introduce new users to setting up their user page, communicating effectively with other users, making basic edits, maintaining a neutral point of view, evaluating content quality, understanding revisions, and using built-in tools like watchlists and history pages to see how articles can be maintained over time. Additional techniques such as adding sources and formatting sections are also introduced in various missions. Each mission consists of a guided tutorial that explains a policy or editing technique in the context of a specific task and presents a simple challenge that tests the user's understanding of the topic. Although the interface prompts the user to go through the missions in order, they can select missions out of order or exit the tutorial at any point.

TWA introduces the basics of communicating and collaborating with other editors early in the tutorial, thus framing Wikipedia as a community of editors, and not just a repository of articles. Throughout the missions, prompts within the tutorial share key facts about the history and philosophy of Wikipedia as a free, global, open knowledge project. This reinforces the idea that Wikipedia is a collective effort driven by volunteer contributions, and attempts to establish a sense that the user is becoming part of an endeavor larger than themselves. In this way, the gamified missions of the tutorial prepare newcomers to identify their contributions in the context of the broader goals of the community.

[4] https://en.wikipedia.org/wiki/Wikipedia:TWA/Story

[5] See: https://en.wikipedia.org/wiki/Wikipedia:Five_pillars



Indented, signed, and notified!

NEW TOOL EARNED: **Communicator Badge**

Figure 3. Badge for completing the mission about communicating with editors.

Badges

The Wikipedia Adventure uses badges to invite newcomers to take on editing Wikipedia not just as a task, but as a part of their identity. Participants who complete all seven missions will earn a total of 15 badges. At the end of each task, the user receives a badge (e.g., Figure 3) and a pop-up message that congratulates them on their accomplishment. The badge titles are designed to reflect new identities the user has taken on (e.g. Copyeditor, Collaborator) as well as competencies around community norms that they have gained (e.g. Verifiability, Neutrality).

These badges are inspired by barnstars, which are awards that Wikipedia editors give each other to acknowledge valued work [31], and userboxes, which are badge-like labels that editors assign to themselves to communicate their achievements, skills, and aspects of their on-wiki identity. Many Wikipedians' user pages contain dozens of badges and barnstars they have accrued during their tenure. When a TWA participant earns a new badge, it is also placed on their user page where it serves a persistent, public reminder that they have been introduced to the values and principles of the community.

Story and Theme

The tutorial is centered around the scenario of creating the article 'Earth' in collaboration with several fictional editors. Earth was selected as subject for its universal relevance, reflecting Wikipedia's mission as a collaborative project that spans geographic and cultural boundaries. Focusing the missions around a single article allows the tutorial to teach a range of technical skills, and also address community norms such as effective communication and adherence to policy while maintaining continuity and verisimilitude.

The visual theme of TWA is galactic exploration, and the tutorial uses graphical elements that are lush, colorful, and whimsical. The tone of the guiding prompts is conversational and humorous in order to create a relaxed and friendly atmosphere in which a new user feels welcome and free to make mistakes.

Deployment

The Wikipedia Adventure was written and developed by a team of volunteers and Wikimedia Foundation staff.[6] It is implemented through the Guided Tours extension (a framework for creating interactive tutorials for MediaWiki), and was deployed in an alpha-test in October 2013 to a group of 50 editors.

After an initial round of testing and debugging, TWA was released on English language Wikipedia in beta in November 2013. The decision to release an English version of the system first was prompted by both the language competencies of the development team, as well as the fact that English Wikipedia is the largest language edition and has the most globally diverse contributor base. However, the system itself can be easily adapted to other language editions in the future through translation of the tutorial text.

STUDY 1: USER SURVEY

After developing The Wikipedia Adventure, we sought feedback and input from new Wikipedia editors who used the tutorial. We conducted a user survey to collect this feedback and evaluate user perceptions of the system design.

Methods

Alongside the initial beta release of TWA, we invited 10,959 editors to use the tutorial via their talk page. We used a large scale deployment so that a diverse group of English Wikipedia users from across the world could participate in the tutorial and provide initial feedback.

We distributed invitations via user talk page messages on a rolling basis to new editors who satisfied the following criteria: they had created their account within the past 24 hours, they had made at least 2 edits, and they had not yet been blocked or received a Level 4 user warning message on their talk page.[7] We distributed the survey invitation to those who used the tutorial after the initial invitation to play the game. The survey invitation consisted of another talk page message with a link to a Qualtrics survey.

Since we sent invitations through talk pages which are visible to the public, it was technically possible for users who were not in our invitation sample to play the game and take the survey. While the likelihood of this is small (new editors see relatively little traffic from other editors to their talk pages), we cannot rule out the possibility that our survey response data may include responses from community members who did not receive the original invitation.

Measures

In keeping with the goals of the system design, the survey examines how survey participants perceived the impact of TWA on their confidence and engagement in editing Wikipedia, whether it communicated effectively, and whether users were

[6] See: https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Adventure

[7] Wikipedia has four levels of warnings given to users suspected of vandalism. Level 4 warnings are the most severe and are reserved for users who commit extreme or frequent vandalism in bad faith.

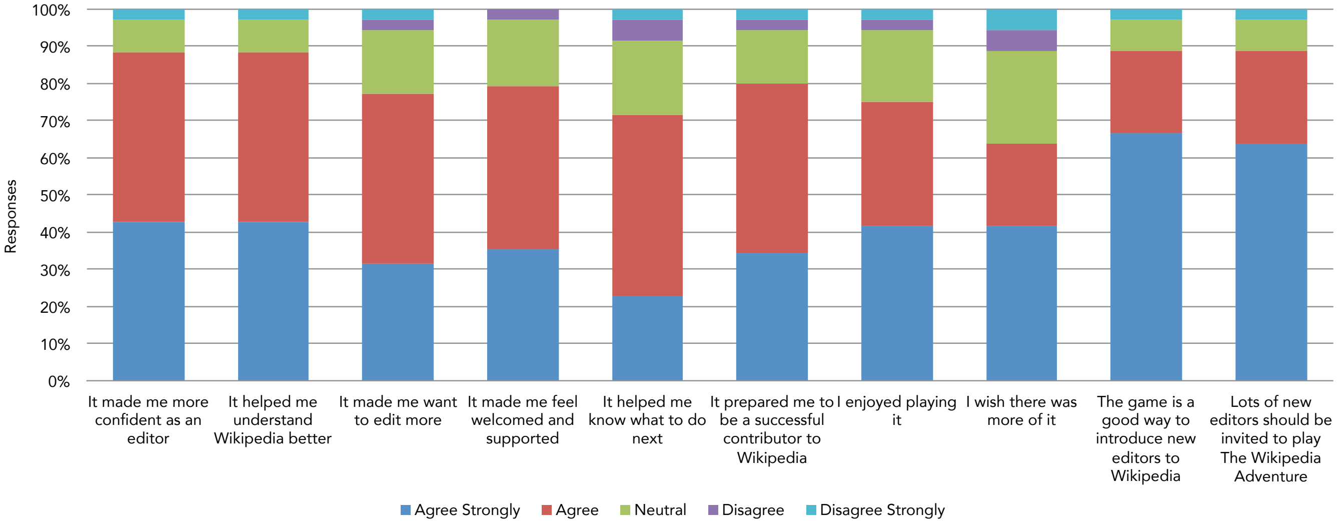


Figure 4. Survey results measuring user confidence, engagement, and satisfaction with The Wikipedia Adventure.

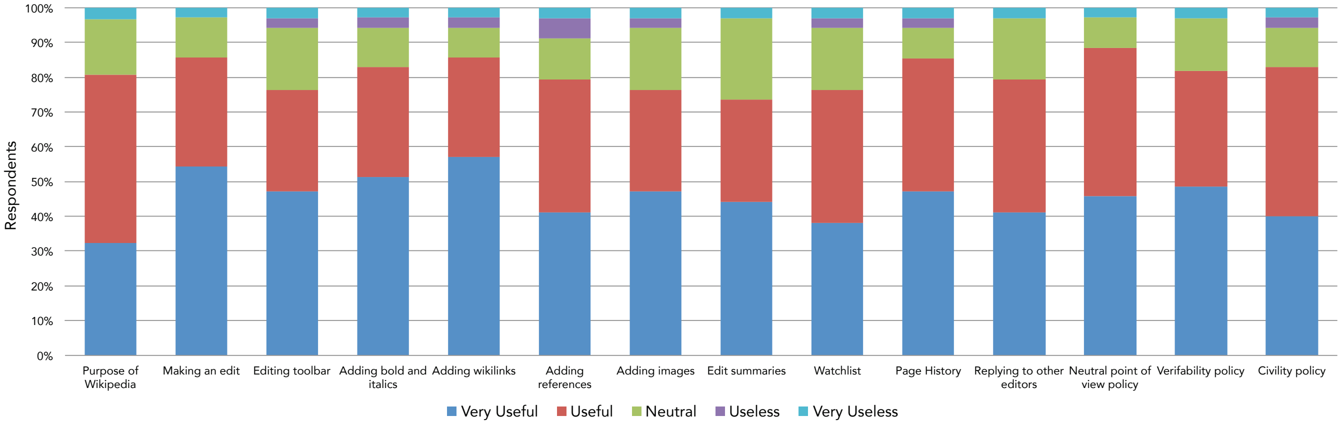


Figure 5. Survey results measuring perceived utility of The Wikipedia Adventure for performing specific editing tasks.

satisfied with the design and tone of the tutorial. We also gathered participants' demographic information.

To measure *user confidence*, we asked users to rate on a 5 point scale the extent to which they agreed or disagreed with statements about TWA. Statements included "It made me more confident as an editor," "It helped me understand Wikipedia better," and, "It prepared me to be a successful contributor to Wikipedia." To measure *user engagement*, we asked users to rate on a 5 point scale the extent to which they agreed or disagreed with statements such as "I enjoyed playing it," "It made me want to edit more," and "I wish there was more of it." To evaluate how clearly TWA communicated information related to editing Wikipedia, we asked respondents to rate on a 5 point scale how well the tutorial taught specific skills such as, "making a wikilink," "adding references," and "adding images." To measure *design satisfaction*, we asked users to rate on a 5 point scale the effectiveness of the interactive elements of the tutorial, their degree of satisfaction with its tone and visual design, and their overall satisfaction with the design.

We also asked participants to select all age groups for which they thought the tutorial would be most appropriate and effective from six choices of age groups that ranged between children below the age of 12, and adults over the age of 55. Finally, we also collected a number of open ended responses from the survey participants.

Results

Out of the 10,959 individuals invited to use the tutorial in the initial beta test, 600 (6%) clicked through and completed at least one mission. From the 600 who used the tutorial and then received an invitation to the survey, 42 individuals (7%) responded between December 23rd, 2013 and January 4th, 2014.

Respondents to the survey came from a number of countries: Australia, Bangladesh, Brazil, Canada, Estonia, Hong Kong, India, Ireland, Nigeria, Portugal, Singapore, Sweden, Macedonia, US, and UK. Although globally diverse, the majority of players came from US or the UK. Close to 11% of the survey respondents identified as female. About 94% had made

100 or fewer edits to Wikipedia, suggesting that our sampling successfully targeted newer editors.

The survey responses suggest that participants felt TWA was an effective way to welcome newcomers and teach them about Wikipedia. We found that 90% of respondents agreed or strongly agreed with the statements "It made me more confident as an editor," and "It helped me understand Wikipedia better,", suggesting that TWA could help build confidence among new editors (see Figure 4). We also found that 80% of respondents agreed or strongly agreed with the statements "It made me want to edit more" and "It made me feel welcomed and supported," which suggests that TWA was helping build engagement among newcomers.

In terms of specific Wikipedia-related skills, over 85% of respondents reported that TWA was "useful" or "very useful" in explaining the neutral point of view policy, how to make an edit, how to add wikilinks, and how to view page histories (see Figure 5). Overall, 91% of respondents found The Wikipedia Adventure "useful" or "very useful" as an introductory tutorial on editing Wikipedia.

Open-ended responses provide additional support for these findings. For example, one respondent noted that "the interactiveness of The Wikipedia Adventure was an easier and better way to learn the basics of Wikipedia versus trying to run around to different pages and just reading about it."

Since The Wikipedia Adventure incorporates a narrative that uses space metaphors and graphics of cartoon aliens, we measured whether these elements appealed to new users. When asked if they would have preferred if TWA had a more 'serious' tone and design, 70% of respondents reported that they liked the tutorial as it was, while 14% wished it were more serious. 76% of participants responded positively to the interactive elements of the tutorial. Respondents suggested that the most appropriate age groups for the tutorial were teens and young adults between the ages of 13 and 22. Overall, 83% of respondents were "satisfied" or "very satisfied" with the design of the tutorial.

One respondent noted:

> It was all beautifully designed. I enjoyed aspects such as the challenges and badges that made it feel more like an educational tool or game rather than a lecture, and [the way that it] recorded your achievement to date.

The positive response to the interactive elements of TWA, seen through both the survey questions as well as the open ended responses, provides validation of our choice to gamify the tutorial. Another respondent said, "I completed the entire game because it wasn't as dry as other training tools out there," suggesting that features such as the interactivity and narrative structure help maintain interest in learning how to edit Wikipedia.

Overall, the survey participants found TWA an effective and useful tool. In particular, respondents valued the general introduction to Wikipedia provided, and reported that the system improved their understanding of Wikipedia and gave them more confidence to edit.

STUDY 2: FIELD EXPERIMENT

As we suggested in our discussion of gamification, the nature of the relationship between gamified participation and participants' subsequent engagement in the task they have learned about remains an active area of research in social computing. Even if a gamified tutorial is engaging, enjoyable, and effective at supporting learning, this might not translate into increased participation after the tutorial is over. Will playing a gamified tutorial like TWA lead to more contributions from newcomers on Wikipedia? The potential for increased enjoyment, confidence, and feelings of self-efficacy are reasons to believe it will. To this end, we use a large-scale, invitation-based field experiment on Wikipedia to evaluate the effect of TWA on the contribution activity of newcomers.

While the survey in Study 1 measured the subjective user experience of playing the game, an experiment tests for causal, behavioral effects of the tutorial on users' subsequent contributions outside of the game's learning environment. A *field* experiment that deploys TWA "in the wild" allows us to estimate the effects that such an intervention would have on Wikipedia newcomers at a large scale.

We use an invitation (or encouragement) design in which some users are randomly invited via their talk pages to play TWA. The Wikipedia community has a tradition of allowing low-barrier contributions on any topic without requiring contributors to do so much as register an account. Due to this tradition and the significant length of the tutorial, the plan for deploying TWA required users to opt-in to the system. In other words, self-selection is a part of the system, and the evaluation of the system's impact needs to incorporate this element of the design to account for the fact that many newcomers might choose to never play the game at all. The choice to use an invitation-based experiment design thus supports the analytic clarity and realism of the experimental results.

TWA was designed to reduce skill-related barriers to entry in editing Wikipedia and to provide an institutionalized, gamified, introduction to concepts like making an edit, using wikimarkup, and communicating via talk pages. Following the evidence of prior work discussed above and consistent with the results from Study 1, we hypothesized that newcomers who played TWA (conditional on having received an invitation to do so) would make:

(H1) *an increased number of contributions overall*,

(H2) *an increased number of contributions on talk pages*,

(H3) *contributions of greater average quality.*

Methods

These hypotheses were tested using an experiment that followed the deployment described in Study 1, after the system was no longer in beta. Starting in February 2014 and continuing over a period of three months, we identified accounts on English Wikipedia to be included in our study on a rolling basis using the same criteria used for inclusion in Study 1. Qualifying accounts were identified on a daily basis and randomly sorted into treatment and control groups. For each user in the treatment group, we sent an invitation to play TWA via their talk page within two days of the creation of their account. We designed the invitations to closely resemble the way that such an intervention might be rolled out on Wikipedia at a large scale. The invitation incorporated graphics from TWA which contrast heavily with other text on Wikipedia, and was thus more likely to be noticed. The invitations were sent out via HostBot (which has been used in the past to invite newcomers to the Teahouse [33]) and logged-in users in the treatment group received a notification that they had been sent an invitation.

To ensure that every participant in our sample had a chance to see their invitation, we only included participants who made at least one edit after getting invited in the analysis. To maintain an equivalent sampling procedure in the control group, we kept only those editors who made at least one edit after the time they would have been invited had they been assigned to receive an invitation. We observe no evidence of crossover between the treatment and control groups—i.e. no users in the control group used TWA.

In total, we identified 1,967 accounts to be included in our study. Of these, 1,751 (89%) were randomly selected to form our treatment group and were invited to play TWA. The other 216 users in our study were placed in a control group and received no invitation. We chose an imbalanced design based on preliminary evidence that the uptake of the invitations to play the game would be low (6% in Study 1). Of the 1,751 users invited, 386 (22%) completed at least some portion of the tutorial. This increase in uptake compared to Study 1 may be due to changes in the invitation text.

We chose to observe the editing behavior of every user over 180 days after their *date of inclusion* in the study. The date of inclusion for a user in the treatment group is either the date that invitation was sent or, for users in our control group, the date the invitation *would have been sent* had the user been in treatment. Although a longer data collection period provides more time to observe systematic variance between the treatment and control groups, it can raise concerns that differences long after the intervention may not be justifiably attributed to the intervention. Our 180 day window was chosen because it is as long as any previous field experiment or system deployment in Wikipedia that we have seen [33, 47].

Measures

Our dependent variables consist of three measures corresponding to each of our hypotheses. To test H1, we measure the overall contributions as the *total number of edits* made by each user in the 180 days after their date of inclusion in the study. This count excludes edits made to the subjects' user pages and user talk pages because TWA automatically generates edits that show up as contributions to these pages. We count all others edits made to Wikipedia including those that were subsequently reverted or deleted. Our results are not substantively affected by the decision to include reverted or deleted edits.

To test H2, we measure the extent to which each subject interacted with others on Wikipedia as the *total number of edits they made to talk pages* on Wikipedia in the 180 days from the time of their inclusion in the study. This variable reflects the emphasis that TWA places on the community dimension of the system.

To test H3, we measure the *average quality of contributions* for each subject by calculating a measure of content "persistence" for all contributions to article pages using metrics developed in parallel by Adler et al. [1, 2] and Halfaker et al. [24, 39]. We estimate the quality of each edit, e_i, by calculating the number of subsequent edits within a fixed radius of subsequent edits in which each word in e_i persists before it is changed or removed. Our measure is the average persistent word score of the article edits made by the user in the 180 days from their inclusion in the study.

Although other radii have been used in research [16], we adopt a radius of 6 edits because this is what is used by Adler et al. [1, 2] in WikiTrust – the most frequently used and widely validated content persistence implementation. Following WikiTrust, we also collapse sequential edits by the same user. Although this will underestimate the productivity of users who edit very infrequently edited articles, we find no statistically significant difference between the mean size of radiuses for edits made by users in our treatment and control groups.

Our key independent variables are two dichotomous measures that indicate whether a particular user was *invited* to play TWA, and whether or not they subsequently *played* it. We consider a user to have played the game if they completed any part of the game, regardless of whether they completed one mission or all seven. Our results are not affected by incorporating the number of missions played, and so we report models that use a dichotomous version of this measure.

Finally, we include a categorical measure capturing the date on which the subjects were incorporated into the research sample. Because random assignment took place within these *sample dates*, this works as a blocking variable that controls for unobserved heterogeneity introduced by running the study over several months [20].

Analytic Approach

Our analysis examined two different facets of the intervention: the effect of inviting a user to play TWA on subsequent contributions, and the effect of playing TWA, conditional on having been invited to do so, on subsequent contributions. Because invitation-based designs are uncommon in social computing research, we explain our analytic approach in detail below.

We estimate the effect that an invitation to use TWA has on a user's subsequent contributions by comparing invited users to users who received no invitation. This is known as an *intent-to-treat* (ITT) estimator, because it does not presuppose that all invited users necessarily received the experimental intervention (i.e. the experience of using the system) [20, 35]. Our ITT models provide unbiased estimates of the effect of distributing the invitations. For each dependent variable, the ITT model (M_{ITT}) takes the generic form:

$$(\mathrm{M_{ITT}}) \quad Y \sim \mathit{invited} + \mathit{sample.date} + \varepsilon$$

To test the impact of playing TWA, we estimate the effect of playing the tutorial conditional on being invited to do so.[8] Experiments of this type, also known as "encouragement designs," are analyzed using two-stage least squares regression (2SLS) [4, 13, 20]. In the first stage of 2SLS we estimate the likelihood that an invitation predicts playing TWA. The vector of fitted values from the first stage model then becomes a predictor in a second stage model which estimates the relationship with our outcome variables. Since the invitation was randomly assigned, the fitted values of the first stage model capture the variation in gameplay caused by the treatment. The second stage model produces an unbiased estimate of the causal effect of playing the game conditional on receiving an invitation.[9] In generic form, the first stage ($\mathrm{M1_{2SLS}}$) and second stage ($\mathrm{M2_{2SLS}}$) models we use for 2SLS are:

$$(\mathrm{M1_{2SLS}}) \quad \mathit{played} \sim \mathit{invited} + \mathit{sample.date} + \varepsilon$$
$$(\mathrm{M2_{2SLS}}) \quad Y \sim \widehat{\mathit{played}} + \mathit{sample.date} + \varepsilon$$

Intent-to-treat and two-stage least squares estimators provide unbiased estimates of treatment effects because of the encouragement design of the study. Unlike a more typical lab experiment or A/B test, we observe relatively low uptake of the game by individuals in the treatment group. Although we might observe that users who played TWA contributed more, on average, than users in the control group (an actual relationship in the data), we must account for the fact that the vast majority of invited (treated) users never visited the tutorial. It is the full treatment and control groups that are "equal in expectation" prior to treatment assignment, and it would be misleading to compare the few users in the treatment group who worked through the tutorial to the full control group who, in large part, would never have done so [35, 42].

Participation is highly skewed in Wikipedia (i.e. a tiny percentage of editors make a large proportion of the total edits) and all of our dependent variables are over-dispersed count measures (see Table 1). As a result, we use negative binomial regression models for all of our estimates. This is typical for highly-skewed count variables and has been applied in prior field experiments on Wikipedia [41]. As a part of our ITT estimate of treatment effects, we also conduct a non-parametric Mann-Whitney U test to identify rare effects by estimating whether the dependent variables for the treatment and control groups are drawn from the same distribution [42]. For all regression models, we report heteroskedasticity and cluster-robust standard errors [4].

Measures	Min.	Median	Max.	Std. Dev.
Total edits	0	6	5282	159.97
Talk Page edits	0	0	365	11.84
Avg. edit quality	0	2.06	6	2.34

Table 1. Summary statistics for dependent variables.

Results

Results from the experiment are shown in Tables 1, 2, 3, and 4 as well as Figure 6. In total, 386 (22%) of those invited completed at least some part of the game. Table 1 describes the distributions of our key measures.

Table 2 shows how many users dropped out of the game after each mission. We find that 181 out of 386 (46%) played till the seventh (and final) mission. The highest dropoff, 93 (24%), occurred after the first mission, with smaller numbers dropping out for most subsequent missions. After the first mission many users kept playing all the way to the final mission.

Mission	Topic	Attrition
1	Editing user page	93
2	Using talk pages	40
3	Editing articles	18
4	Neutral point of view	9
5	Verifying sources	11
6	Civil discussion	34
7	Adding sections	181

Table 2. The attrition for every mission is measured as the number of subjects who play some part of the mission but did not go on to play subsequent missions. For example, 93 subjects played some part of Mission 1, but did not proceed to Mission 2. A total of 386 subjects played TWA.

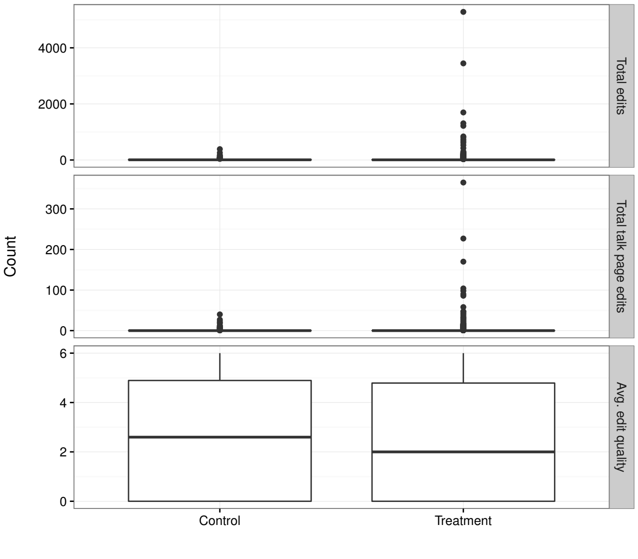


Figure 6. Boxplots showing the distributions of outcomes for total edits (top), talk page edits (middle), and average edit quality (bottom) across subjects assigned to the treatment and control conditions.

Figure 6 plots the distribution of all three outcome measures across treatment and control conditions. We note that for the first two outcomes (total edits and total talk page edits) many

[8] Note that subjects who played TWA are a (small) subset of subjects who were invited to do so. As a result, comparing the outcomes for the set of TWA players against the entire control group cannot recover an unbiased estimate of the effect of playing the game. The control group contains subjects who would have played and those who would have not (had they been invited).

[9] We refer interested readers to several key references for formal details and proofs of 2SLS [3, 13, 20, 35]. We are not familiar with prior work in social computing that applies these methods.

Outcome	Diff. of medians	Test statistic	p-value
Total edits	-1	173492.5	0.05
Talk Page edits	0	190527	0.79
Avg. edit quality	-0.6	181712.5	0.34

Table 3. Results of non-parametric Mann-Whitney U tests for all dependent variables.

subjects in both the treatment and control groups registered no further edits of either kind following the intervention. In terms of the distributions, we see that the treatment condition had a longer tail (a handful of extremely high outcomes) along both of these measures. The third boxplot illustrates that outcomes for average edit quality are distributed in a nearly identical fashion across the two conditions.

We report the results of non-parametric Mann-Whitney U tests in Table 3. The table includes the difference in medians ($\mu_{Y|treatment} - \mu_{Y|control}$), the value of the test statistic U, and the corresponding p-value for each dependent variable. The results indicate that the distributions of talk page edits and average edit quality are not different. In the case of total edits, the difference of medians is -1 and the p-value reaches conventional levels of significance, suggesting evidence of a negative distribution shift caused by the treatment. We interpret the results of the test as weak evidence of a statistically meaningful variation between the individuals invited to play TWA and those who were not. We suspect that this might be because the time taken by new editors to play TWA cut into the potential time they had to contribute to Wikipedia early on.

In our intent-to-treat analysis using negative binomial regression to estimate the effect of an invitation to TWA, our models produce small estimates for all our dependent variables that are not statistically distinguishable from zero (see the top part of Table 4).

When we test the effect of playing TWA with 2SLS (see the bottom part of Table 4), we also find null effects with small coefficients for all dependent variables. For all coefficients, the standard errors are relatively large compared with the estimates and none approach conventional levels of statistical significance.

The parameter estimates for our null results represent well-estimated zeroes and suggest that any underlying effect we are unable to estimate with our sample would likely be extremely small. Post-hoc power analysis shows that if a data set of this size displayed even a small effect size (0.2 standard deviations), we would have had a 99% chance of detecting it at the 0.05 significance level. Thus, we can conclude that TWA does not alter the quantity or quality of newcomers' contributions to Wikipedia.

As a robustness check, we also estimated models using measures of our dependent variables computed over both 360 and 60 days following inclusion in the study. We again find null results for all three hypotheses over 360 days as well as for H2 and H3 over 60 days. Echoing our Mann-Whitney U test, we find a small negative relationship in our test for H1 in the 60 day dataset. In these results, we estimate that invitation to the game (ITT) was associated with approximately 1.25 fewer edits and approximately 3 fewer edits in our 2SLS model over the 60 days after inclusion. One potential explanation is that participation in the tutorial may have supplanted other editing among participants in the treatment group but that this effect is "washed out" over time. In any case, the pattern of results across the three study lengths is not consistent with predictions from previous work that the system would cause new users to contribute more. If anything, there is weak evidence suggesting that TWA might have caused them to make several fewer edits in the period immediately following inclusion in the study.

DISCUSSION

The results from Study 1 show that respondents found TWA to be a useful and satisfying tool for learning how to edit Wikipedia. In particular, study participants valued having a system that provided a general introduction to Wikipedia, and stated that it improved their understanding of the community and gave them more confidence to edit. Study 2, however, shows that despite the perceived effectiveness of the design and the satisfaction of the users, playing TWA did not alter the subsequent behavior of newcomers on Wikipedia.

The survey responses validated the idea of using gamification to introduce an institutionalized form of socialization to Wikipedia. Users found the gamified aspects of the tutorial rewarding and engaging and agreed that a tutorial that provides a broad overview of editing should be shared with new editors on Wikipedia. These findings suggest that we accomplished our system design goals and that the tutorial provided a compelling and enjoyable institutionalized introduction to the skills, norms, and expectations involved in becoming a Wikipedian. We believe these findings validate some of the claims of prior gamification research as well the theoretical justification for pursuing institutionalized socialization as a complement to existing onboarding systems in Wikipedia.

Study 1 has several limitations, including the small pool of respondents and the fact that the survey can not capture the reasons why some individuals chose not to play the game. The limited uptake of the game and low response rate of the survey mean that our Study 1 findings might not extend to all the individuals invited to play TWA or even to all users who played it as part the initial deployment. Additionally, the survey does not assess whether respondents who played TWA effectively learned the skills that the tutorial sought to introduce. Finally, the wording of the survey questions could have elicited overly positive responses due to satisficing behavior and social desirability pressures. Nevertheless, the results of Study 1 made us optimistic that some newcomers would elect to play TWA and that doing so could have a positive impact on their subsequent engagement through increased enjoyment in learning how to edit Wikipedia and improved confidence and self-efficacy.

However, contrary to predictions from organizational and social computing theory and design, Study 2 shows that TWA had no measurable impact on newcomer participation. All of our statistical tests for regressions of all three outcome measures fail to reject the null hypothesis of no effects regard-

Estimand	Model	Dependent Variable		
		Total edits	Talk page edits	Avg. quality
Invited to play TWA	Negative binomial	-0.107	-0.146	-0.065
		(0.112)	(0.273)	(0.088)
Playing TWA conditional on invitation	2SLS	-0.545	-0.730	-0.155
		(0.478)	(1.151)	(0.375)

Table 4. Regression results estimating the effects of (1) the invitation to play TWA and (2) playing TWA conditional on having been invited to do so on three measures of newcomer participation. For each dependent variable, we provide coefficients with standard errors in parentheses. The models reported here all include a control (unreported) for the number of days that each participant had edited Wikipedia. The results are substantively unchanged when we drop the control.

less of the model specification or the estimand (ITT or playing TWA conditional on receiving an invitation). Robustness checks conducted on a smaller data collection window suggest that TWA may have even reduced contribution rates in the short term. We conclude that these results demonstrate a null effect.

The limitations of Study 2 include the possibility that the invitation process or deployment of the system might have shaped the outcomes in ways we cannot detect. The visual differences between the outcome distributions in Figure 6 also suggest that we might explore alternative, novel estimation techniques focused on detecting rare, large effects [42]. It is also possible that the treatment impacted some other outcome variable that we do not measure in this study, such as survival rate. Future research might explore these questions.

What factors might explain the null effects in Study 2? The statistical tests we report are appropriate to the design and the study had adequate statistical power to detect treatment effects had they existed. We propose several interpretations, which focus on the system design, the culture of the Wikipedia community, the self-selected and voluntary nature of participation in peer production communities, and the limitations of gamified interactive systems.

System design factors

Shortcomings in the design of TWA offer one possible explanation of the null effects in Study 2. If the tutorial itself was poorly designed, a better implementation may have altered new editor contributions. For example, if TWA users perceived editing within a sandboxed environment instead of a live Wikipedia page as an illegitimate or inauthentic form of participation [32], this might have undermined the system's effects. Although other work has suggested that designs like ours can overcome these effects [21, 29, 43], it remains a possible explanation and a design choice worth revisiting.

The fact that the survey respondents in Study 1 overwhelmingly perceived TWA to be positive and well-designed also suggests that the system design did not have glaring shortcomings. Subsequent to the completion of this study, many more TWA users have also given the tutorial glowing feedback through the comment box on the game's webpage. This implies that limitations of the particular system design do not fully explain the null result.

The temptation to attribute shortcomings visible only after field testing to details of our implementation of TWA points to a larger concern about understanding the impact of any new system. A creative and thoughtful designer can always imagine alternative approaches that *might* transform the effects of an existing system, no matter how carefully planned or executed the existing design may be. Our findings in these two studies indicated that TWA's system design satisfied the criteria and goals of the system's creators, as well as the system's early users. Even so, it did not produce the effects predicted by either theory or preliminary testing. For this reason, we believe limitations in the existing theories as well as the specific conditions of TWA's deployment better explain the observed outcomes.

Cultural factors related to Wikipedia

Prior research points to several ways that Wikipedia's existing culture may have undermined TWA's expected effects. The new editors in our study may have had unpleasant experiences during their initial time on Wikipedia that negated any potential motivational benefits they may have gained from playing TWA. Even for experienced contributors, the abrasive and hostile tone of interactions among Wikipedia editors deters participation.[10] Many new editors receive multiple warning messages on their talk pages within their first few editing sessions [19]. These warning messages, the majority of which are automated, strongly-worded, and accompanied by a revert, can drive new editors away from the site [22].

Just observing toxic exchanges among other Wikipedians could have convinced the new editors in our study that the Wikipedia community is not a welcoming place. A lighthearted, automated tutorial depicting a collegial collaboration process may not be sufficiently compelling to counteract these negative observations or experiences. We cannot confirm or reject this possibility fully through our empirical analysis because we do not know what perceptions study participants who dropped out of editing may have had.

Limitations of gamification

Most previous studies of gamified systems have focused on subjective measures of engagement with, and enjoyment of, the system. With few exceptions, they have not evaluated the impact of gamified systems on subsequent performance. Our study is one of the first to assess the effect of a gamified learning system on engagement behaviors outside of the system itself. Although some studies have shown that gamification can support learning, meta-analyses have suggested that the few studies that did analyze impact on performance did not

[10] This is based on the Wikimedia Foundation's survey of former contributors referred to in a previous footnote.

reliably show improvements [26]. This may help explain the contrast between our survey results and our field experiment.

Another explanation stems from the shift in incentives between the game and the "real" world of Wikipedia contributing. It is possible that the extrinsic motivation provided by the gamified tutorial was simply not replaced by intrinsic motivations needed to drive subsequent contributions. In a project like Wikipedia that depends heavily on intrinsically motivated members to make contributions, a gamified tutorial may be helpful and fun to use, but ultimately unsuccessful at building long-term commitment and retention. This echoes our earlier point about the possible limitations of sandboxed learning environments. The current study contributes to our understanding of the effectiveness of gamification by presenting both a subjective evaluation of a novel gamified system, as well as a measurement of its subsequent impact in a non-gamified context.

Self-selection and voluntary participation

The voluntary nature of participation and membership in peer production communities like Wikipedia offers another possible explanation. Contributors engaged in peer production self-select into their preferred communities, tasks, and social roles [6, 7, 46]. Institutionalized training programs may be more effective in more formal organizations because newcomers cannot select out of them as easily. TWA provides a general overview of contributing to Wikipedia, but a person who is interested solely in learning how to accomplish one specific task (e.g., fixing a citation) might find the full tutorial burdensome and choose not to play. An institutionalized socialization approach might also fail because it tries to artificially speed up the process of "becoming Wikipedian", which involves a gradual transformation from participating peripherally to seeing oneself as part of the community [8].

The dynamics of self-selection may best explain the null effect of Study 2. Because TWA depends on users to choose to play the game, those who do so are likely to vary systematically from those who do not. Specifically, the newcomers who received and accepted our invitation may be more motivated, committed, or skilled than those who received the invitation and chose not to play. Playing the game may have given the exceptional newcomers a positive experience without impacting the quantity or quality of their subsequent contributions. The fact that a subset of individuals in both the treatment and control conditions went on to make numerous edits of high quality supports this idea. Our analytic approach with 2SLS supports this inference in that we identify the causal effect of playing TWA conditional on having received the invitation. The null findings in these models indicate that the people who played the game and went on to contribute extensively would have done so anyway. We cannot say more generally whether Wikipedians may be "born," "made," or some combination of the two [28, 37]. We conclude that TWA did not make active editors out of people who would have been inactive in the absence of the game.

This study illustrates the value in evaluating novel systems in "live" field deployments within communities. As discussed above, the findings of Study 1 validate many of the design principles and findings from prior literature on gamification and newcomer socialization. However, Study 2 revealed that these principles and findings cannot explain the empirical impact of the system. Study 2 does not invalidate prior work on institutionalized socialization or gamification, but it does show that successful newcomer orientation in a volunteer community like Wikipedia remains a compelling design challenge.

While the invitation-based field deployment of TWA yielded no effect on newcomer contributions, it is possible that such a system would work better in contexts where newcomers are required to play it before editing Wikipedia, thus circumventing the self-selection issue. For instance, using the tutorial in a classroom setting where students are required to contribute to Wikipedia (increasingly common through initiatives such as the Wikipedia Education Program) might produce positive results.

CONCLUSION

We designed and evaluated The Wikipedia Adventure, a gamified, interactive tutorial that extended techniques of institutionalized socialization to newcomers in Wikipedia. The first part of our evaluation, a user survey, validated the principles, theories, and goals of TWA's design. The second part of our evaluation, an invitation-based field experiment, revealed that deploying the system did not alter newcomer contribution patterns over several months. We suggest that the null findings may be due to a combination of factors including the culture of Wikipedia, limitations of gamified systems, and the dynamics of self-selection in voluntary peer production communities. We believe that our second study represents the first invitation-based field experiment and application of two-stage least squares in the social computing literature.

Given the positive results in Study 1 and the strong theoretical support for its design, we were genuinely surprised by the null result in Study 2. The discrepancies between the results in our two studies point to an important secondary contribution of our work. Despite the positive response from users that were surveyed in Study 1, our field experiment in Study 2 demonstrated clearly that subjective perceptions of utility and usability do not necessarily translate into lasting changes in user behavior.

Although null results can be difficult to convincingly establish and interpret, they can play an important role in contributing to our knowledge of social computing theories and systems. TWA's design was informed by previous empirical, theoretical, and systems work, and it performed well according to the types of survey self-report measures used to evaluate the usability of many social computing systems. Post-hoc power analysis suggests that our estimates in Study 2 are well estimated zeros and that our sample size is sufficiently large to detect even small effects. Our work does not provide the final word on institutionalized socialization or gamified tutorials in peer production. That said, we believe it contributes to our understanding of these topics through both what we have been able to show, as well as what we have not.

Finally, we believe that our work shows how any intervention that attempts to assimilate new users into an existing peer production community might be limited when deployed in the wild. Wikipedia has complex norms and rules for participation which are obscure to newcomers. Institutionalized and gamified socialization systems like TWA may inform the design of future orientation systems, but more profound changes to the interface or modes of interaction between editors might also be needed to increase contributions from the targeted groups.

ACKNOWLEDGEMENTS

The authors would like to thank our CSCW reviewers and Associate Chairs for their thoughtful and detailed feedback, which greatly improved this paper. We would also like to thank the Wikimedia Foundation for its assistance with the data collection process, as well as the BYOR Workshop at Northwestern, the Community Data Science Collective, and Marlon Twyman for feedback. Portions of the project were completed at the Helen R. Whitely Retreat Center at the University of Washington's Friday Harbor Laboratories and using the Hyak high performance computing cluster at the University of Washington. Financial support for this work came from the National Science Foundation (grants IIS-1617129 and IIS-1617468), Northwestern University, and the University of Washington.

ACCESS TO DATA

A replication dataset has been placed in the Harvard Dataverse archive and is available at the following URL: http://dx.doi.org/10.7910/DVN/6HPRIG.

REFERENCES

1. B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning Trust to Wikipedia Content. In *Proceedings of the 4th International Symposium on Wikis (WikiSym '08)*. ACM, New York, NY, USA, 26:1–26:12. DOI: http://dx.doi.org/10.1145/1822258.1822293
2. B. Thomas Adler and Luca de Alfaro. 2007. A Content-driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 261–270. DOI: http://dx.doi.org/10.1145/1242572.1242608
3. Joshua D Angrist and Alan B Krueger. 1990. *Does compulsory school attendance affect schooling and earnings?* Technical Report. National Bureau of Economic Research.
4. Joshua D. Angrist and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.
5. Talya N Bauer, Todd Bodner, Berrin Erdogan, Donald M Truxillo, and Jennifer S Tucker. 2007. Newcomer adjustment during organizational socialization: a meta-analytic review of antecedents, outcomes, and methods. *Journal of applied psychology* 92, 3 (2007), 707.
6. Yochai Benkler. 2002. Coase's penguin, or, Linux and the nature of the firm. *Yale Law Journal* 112, 3 (2002), 369–446. http://yalelawjournal.org/112/3/369_yochai_benkler.html
7. Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer Production: A Form of Collective Intelligence. In *The Handbook of Collective Intelligence*, Michael Bernstein and Thomas Malone (Eds.). MIT Press, Cambridge, MA.
8. Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP '05)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/1099203.1099205
9. Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't Look Now, but We'Ve Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1101–1110. DOI: http://dx.doi.org/10.1145/1357054.1357227
10. Giovanni Luca Ciampaglia and Dario Taraborelli. 2015. MoodBar: Increasing New User Retention in Wikipedia Through Lightweight Socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 734–742. DOI: http://dx.doi.org/10.1145/2675133.2675181
11. Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 383–392. DOI: http://dx.doi.org/10.1145/2145204.2145265
12. Edward L. Deci and Richard M. Ryan. 1975. Intrinsic Motivation. In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc., New York NY. http://onlinelibrary.wiley.com/doi/10.1002/9780470479216.corpsy0467/abstract
13. Thomas S Dee. 2004. Are there civic returns to education? *Journal of Public Economics* 88, 9 (2004), 1697–1720.
14. Sylvain Dejean and Nicolas Jullien. 2015. Big from the beginning: Assessing online contributors' behavior by their first contribution. *Research Policy* 44, 6 (July 2015), 1226–1239. DOI: http://dx.doi.org/10.1016/j.respol.2015.03.001
15. Nicolas Ducheneaut. 2005. Socialization in an Open Source Software Community: A Socio-Technical Analysis. *Computer Supported Cooperative Work (CSCW)* 14, 4 (2005), 323–368. DOI: http://dx.doi.org/10.1007/s10606-005-9000-1

16. Michael D. Ekstrand and John T. Riedl. 2009. rv you're dumb: Identifying Discarded Work in Wiki Article History. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*. ACM, New York, NY, USA, 4:1–4:10. DOI: http://dx.doi.org/10.1145/1641309.1641317

17. Rosta Farzan, Robert Kraut, Aditya Pal, and Joseph Konstan. 2012. Socializing Volunteers in an Online Community: A Field Experiment. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 325–334. DOI: http://dx.doi.org/10.1145/2145204.2145256

18. Bruno S. Frey and Reto Jegen. 2001. Motivation Crowding Theory. *Journal of Economic Surveys* 15, 5 (2001), 589–611. DOI: http://dx.doi.org/10.1111/1467-6419.00150

19. R. Stuart Geiger, Aaron Halfaker, Maryana Pinchuk, and Steven Walling. 2012. Defense Mechanism or Socialization Tactic? Improving Wikipedia's Notifications to Rejected Contributors. In *Sixth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, Dublin, Ireland, 122–129. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4657

20. Alan S Gerber and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton, New York.

21. Mark Guzdial and Allison Elliott Tew. 2006. Imagineering Inauthentic Legitimate Peripheral Participation: An Instructional Design Approach for Motivating Computing Education. In *Proceedings of the Second International Workshop on Computing Education Research (ICER '06)*. ACM, New York, NY, USA, 51–58. DOI: http://dx.doi.org/10.1145/1151588.1151597

22. Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013a. The Rise and Decline of an Open Collaboration System How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (May 2013), 664–688. DOI: http://dx.doi.org/10.1177/0002764212469365

23. Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. 2013b. Making peripheral participation legitimate: reader engagement experiments in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*. ACM, New York, NY, USA, 849–860. DOI: http://dx.doi.org/10.1145/2441776.2441872

24. Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*. ACM, New York, NY, USA, 15:1–15:10. DOI: http://dx.doi.org/10.1145/1641309.1641332

25. Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, USA, 163–172. DOI: http://dx.doi.org/10.1145/2038558.2038585

26. Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work? - A Literature Review of Empirical Studies on Gamification. In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences (HICSS '14)*. IEEE Computer Society, Washington, DC, USA, 3025–3034. DOI: http://dx.doi.org/10.1109/HICSS.2014.377

27. Michael D. Hanus and Jesse Fox. 2015. Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education* 80 (Jan. 2015), 152–161. DOI: http://dx.doi.org/10.1016/j.compedu.2014.08.019

28. Shih-Wen Huang, Minhyang (Mia) Suh, Benjamin Mako Hill, and Gary Hsieh. 2015. How Activists Are Both Born and Made: An Analysis of Users on Change.Org. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 211–220. DOI: http://dx.doi.org/10.1145/2702123.2702559

29. Diana Joseph and Denise C. Nacu. 2003. Designing Interesting Learning Environments When the Medium isn't enough. *Convergence: The International Journal of Research into New Media Technologies* 9, 2 (June 2003), 84–115. DOI: http://dx.doi.org/10.1177/135485650300900207

30. Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press, Cambridge, MA.

31. Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW2008)*. ACM, San Diego, CA, USA, 47–56. DOI: http://dx.doi.org/10.1145/1460563.1460573

32. Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK.

33. Jonathan T. Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*. ACM, New York, NY, USA, 839–848. DOI: http://dx.doi.org/10.1145/2441776.2441871

34. Gabriel Mugar, Carsten Østerlund, Katie DeVries Hassman, Kevin Crowston, and Corey Brian Jackson. 2014. Planet Hunters and Seafloor Explorers: Legitimate Peripheral Participation Through Practice Proxies in Online Citizen Science. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 109–119. DOI: http://dx.doi.org/10.1145/2531602.2531721

35. Richard J. Murnane and John B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, Oxford ; New York.

36. David R. Musicant, Yuqing Ren, James A. Johnson, and John Riedl. 2011. Mentoring in Wikipedia: A Clash of Cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, USA, 173–182. DOI: http://dx.doi.org/10.1145/2038558.2038586

37. Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work (GROUP '09)*. ACM, New York, NY, USA, 51–60. DOI: http://dx.doi.org/10.1145/1531674.1531682

38. Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Number 4956 in Lecture Notes in Computer Science. Springer, Berlin, Germany, 663–668. http://link.springer.com/chapter/10.1007/978-3-540-78646-7_75 DOI: 10.1007/978-3-540-78646-7_75.

39. Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, Destroying, and Restoring Value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*. ACM, New York, NY, USA, 259–268. DOI: http://dx.doi.org/10.1145/1316624.1316663

40. Joseph Reagle. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. MIT Press, Cambridge Mass.

41. Michael Restivo and Arnout van de Rijt. 2014. No praise without effort: experimental evidence on how rewards affect Wikipedia's contributor community. *Information, Communication & Society* 17, 4 (2014), 1–12. DOI: http://dx.doi.org/10.1080/1369118X.2014.888459

42. Paul R. Rosenbaum. 2010. *Design of Observational Studies*. Springer, New York.

43. David Williamson Shaffer and Mitchel Resnick. 1999. ”Thick” Authenticity: New Media and Authentic Learning. *J. Interact. Learn. Res.* 10, 2 (Dec. 1999), 195–215. http://dl.acm.org/citation.cfm?id=325370.325387

44. Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. The Singularity is Not Near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09)*, Vol. 8. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/1641309.1641322

45. John Van Maanen and Schein, Edgar H. 1979. Toward a theory of organizational socialization. In *Research in organizational behavior*, Barry M. Staw (Ed.). JAI Press, Greenwich, CT, 209–264.

46. Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. 2011. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference (iConference '11)*. ACM, New York, NY, USA, 122–129. DOI: http://dx.doi.org/10.1145/1940761.1940778

47. Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012a. Effectiveness of Shared Leadership in Online Communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 407–416. DOI: http://dx.doi.org/10.1145/2145204.2145269

48. Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012b. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 935–944. DOI: http://dx.doi.org/10.1145/2145204.2145344