Wiki Surveys: Open and Quantifiable Social Data Collection

Matthew J. Salganik[1], Karen E. C. Levy[2]

1 Department of Sociology, Center for Information Technology Policy, and Office of Population Research, Princeton University, Princeton, NJ, USA, **2** Information Law Institute and Department of Media, Culture, and Communication, New York University, New York, NY, USA and Data & Society Research Institute, New York, NY, USA

* mjs3@princeton.edu

Abstract

In the social sciences, there is a longstanding tension between data collection methods that facilitate quantification and those that are open to unanticipated information. Advances in technology now enable new, hybrid methods that combine some of the benefits of both approaches. Drawing inspiration from online information aggregation systems like Wikipedia and from traditional survey research, we propose a new class of research instruments called *wiki surveys*. Just as Wikipedia evolves over time based on contributions from participants, we envision an evolving survey driven by contributions from respondents. We develop three general principles that underlie wiki surveys: they should be greedy, collaborative, and adaptive. Building on these principles, we develop methods for data collection and data analysis for one type of wiki survey, a pairwise wiki survey. Using two proof-of-concept case studies involving our free and open-source website www.allourideas.org, we show that pairwise wiki surveys can yield insights that would be difficult to obtain with other methods.

Introduction

In the social sciences, there is a longstanding tension between data collection methods that facilitate quantification and those that are open to unanticipated information. For example, one can contrast a traditional public opinion survey based on a series of pre-written questions and answers with an interview in which respondents are free to speak in their own words. The tension between these approaches derives, in part, from the strengths of each: open approaches (e.g., interviews) enable us to learn new and unexpected information, while closed approaches (e.g., surveys) tend to be more cost-effective and easier to analyze. Fortunately, advances in technology now enable new, hybrid approaches that combine the benefits of each. Drawing inspiration both from online information aggregation systems like Wikipedia and from traditional survey research, we propose a new class of research instruments called *wiki surveys*. Just as Wikipedia grows and improves over time based on contributions from participants, we envision an evolving survey driven by contributions from respondents.

Although the tension between open and closed approaches to data collection is currently most evident in disagreements between proponents of quantitative and qualitative methods, the trade-off between open and closed survey questions was also particularly contentious in the early days of survey research [1–3]. Although closed survey questions, in which respondents choose from a series of pre-written answer choices, have come to dominate the field, this is not because they have been proven superior for measurement. Rather, the dominance of closed questions is largely based on practical considerations: having a fixed set of responses dramatically simplifies data analysis [4].

The dominance of closed questions, however, has led to some missed opportunities, as open approaches may provide insights that closed methods cannot [4–8]. For example, in one study, researchers conducted a split-ballot test of an open and closed form of a question about what people value in jobs [5]. When asked in closed form, virtually all respondents provided one of the five researcher-created answer choices. But, when asked in open form, nearly 60% of respondents provided a new answer that fell outside the original five choices. In some situations, these unanticipated answers can be the most valuable, but they are not easily collected with closed questions. Because respondents tend to confine their responses to the choices offered [9], researchers who construct all the possible choices necessarily constrain what can be learned.

Projects that depend on crowdsourcing and user-generated content, such as Wikipedia, suggest an alternative approach. What if a survey could be constructed by respondents themselves? Such a survey could produce clear, quantifiable results at a reasonable cost, while minimizing the degree to which researchers must impose their pre-existing knowledge and biases on the data collection process. We see wiki surveys as an initial step toward this possibility.

Wiki surveys are intended to serve as a complement to, not a replacement for, traditional closed and open methods. In some settings, traditional methods will be preferable, but in others we expect that wiki surveys may produce new insights. The field of survey research has always evolved in response to new opportunities created by changes in technology and society [10–16], and we see this research as part of that longstanding evolution.

In this paper, we develop three general principles that underlie wiki surveys: they should be greedy, collaborative, and adaptive. Building on these principles, we develop methods for data collection and data analysis for one type of wiki survey, a pairwise wiki survey. Using two proof-of-concept case studies involving our free and open-source website www.allourideas.org, we show that pairwise wiki surveys can yield insights that would be difficult to obtain with other methods. The paper concludes with a discussion of the limitations of this work and possibilities for future research.

Wiki surveys

Online information aggregation projects, of which Wikipedia is an exemplar, can inspire new directions in survey research. These projects, which are built from crowdsourced, user-generated content, tend to share certain properties that are not characteristic of traditional surveys [17–20]. These properties guide our development of wiki surveys. In particular, we propose that wiki surveys should follow three general principles: they should be *greedy*, *collaborative*, and *adaptive*.

Greediness

Traditional surveys attempt to collect a fixed amount of information from each respondent; respondents who want to contribute less than one questionnaire's worth of information are considered problematic, and respondents who want to contribute more are prohibited from doing
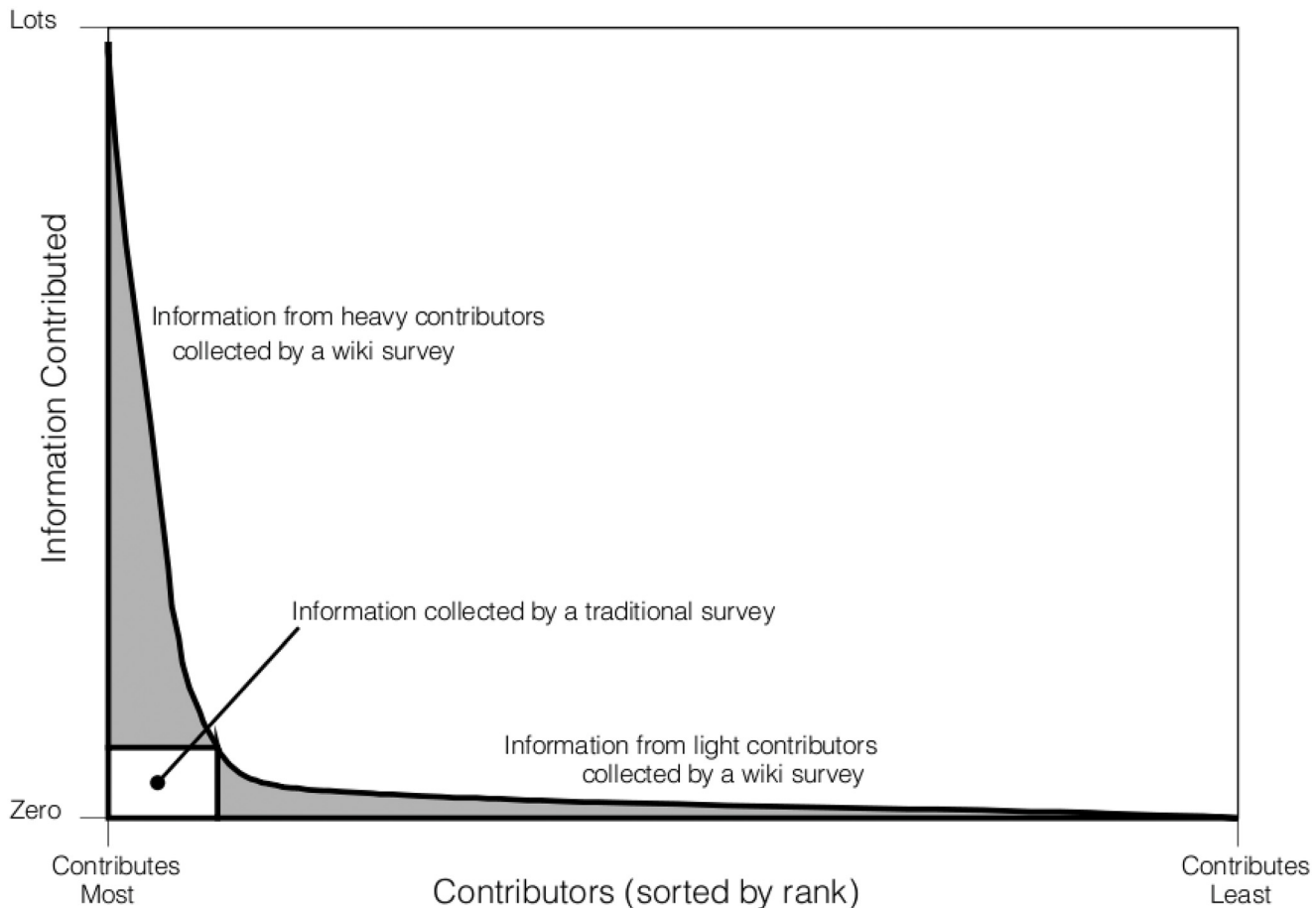
Fig 1. Schematic of rank order plot of contributions to successful online information aggregation projects. These systems can handle both heavy contributors ("the fat head"), shown on the left side of the plot, and light contributors ("the long tail"), shows on the right side of the plot. Traditional survey methods utilize information from neither the "fat head" nor the "long tail" and thus leave huge amounts of information uncollected.

doi:10.1371/journal.pone.0123483.g001

so. This contrasts sharply with successful information aggregation projects on the Internet, which collect as much or as little information as each participant is willing to provide. Such a structure typically results in highly unequal levels of contribution: when contributors are plotted in rank order, the distributions tend to show a small number of heavy contributors—the "fat head"—and a large number of light contributors—the "long tail" [21, 22] (Fig 1). For example, the number of edits to Wikipedia per editor roughly follows a power-law distribution with an exponent 2 [22]. If Wikipedia were to allow 10 and only 10 edits per editor—akin to a survey that requires respondents to complete one and only one form—it would exclude about 95% of the edits contributed. As such, traditional surveys potentially leave enormous amounts of information from the "fat head" and "long tail" uncollected. Wiki surveys, then, should be *greedy* in the sense that they should capture as much or as little information as a respondent is willing to provide.

Collaborativeness

In traditional surveys, the questions and answer choices are typically written by researchers rather than respondents. In contrast, wiki surveys should be *collaborative* in that they are open

to new information contributed directly by respondents that may not have been anticipated by the researcher, as often happens during an interview. Crucially, unlike a traditional "other" box in a survey, this new information would then be presented to future respondents for evaluation. In this way, a wiki survey bears some resemblance to a focus group in which participants can respond to the contributions of others [23, 24]. Thus, just as a community collaboratively writes and edits Wikipedia, the content of a wiki survey should be partially created by its respondents. This approach to collaborative survey construction resembles some forms of survey pre-testing [25]. However, rather than thinking of pre-testing as a phase distinct from the actual data collection, in wiki surveys the collaboration process continues throughout data collection.

Adaptivity

Traditional surveys are static: survey questions, their order, and their possible answers are determined before data collection begins and do not evolve as more is learned about the parameters of interest. This static approach, while easier to implement, does not maximize the amount that can be learned from each respondent. Wiki surveys, therefore, should be *adaptive* in the sense that the instrument is continually optimized to elicit the most useful information, given what is already known. In other words, while collaborativeness involves being open to new information, adaptivity involves using the information that has already been gathered more efficiently. In the context of wiki surveys, adaptivity is particularly important given that respondents can provide different amounts of information (due to greediness) and that some answer choices are newer than others (due to collaborativeness). Like greediness and collaborativeness, adaptivity increases the complexity of data analysis. However, research in related areas [26–33] suggests that gains in efficiency from adaptivity can more than offset the cost of added complexity.

Pairwise Wiki Surveys

Building on previous work [34–40], we operationalize these three principles into what we call a *pairwise wiki survey*. A pairwise wiki survey consists of a single question with many possible answer items. Respondents can participate in a pairwise wiki survey in two ways: first, they can make pairwise comparisons between items (i.e., respondents vote between item A and item B), and second, they can add new items that are then presented to future respondents.

Pairwise comparison, which has a long history in the social sciences [41], is an ideal question format for wiki surveys because it is amenable to the three criteria described above. Pairwise comparison can be *greedy* because the instrument can easily present as many (or as few) prompts as each respondent is willing to answer. New items contributed by respondents can easily be integrated into the choice sets of future respondents, enabling the instrument to be *collaborative*. Finally, pairwise comparison can be *adaptive* because the pairs to be presented can be selected to maximize learning given previous responses. These properties exist because pairwise comparisons are both granular and modular; that is, the unit of contribution is small and can be readily aggregated [17].

Pairwise comparison also has several practical benefits. First, pairwise comparison makes manipulation, or "gaming," of results difficult because respondents cannot choose which pairs they will see; instead, this choice is made by the instrument. Thus, when there is a large number of possible items, a respondent would have to respond many times in order to be presented with the item that she wishes to "vote up" (or "vote down") [42]. Second, pairwise comparison requires respondents to prioritize items—that is, because the respondent must select one of two discrete answer choices from each pair, she is prevented from simply saying that she likes

(or dislikes) every option equally strongly. This feature is particularly valuable in policy and planning contexts, in which finite resources make prioritization of ideas necessary. Finally, responding to a series of pairwise comparisons is reasonably enjoyable, a common characteristic of many successful web-based social research projects [43, 44].

Data collection

In order to collect pairwise wiki survey data, we created the free and open-source website All Our Ideas (www.allourideas.org), which enables anyone to create their own pairwise wiki survey. To date, about 6,000 pairwise wiki surveys have been created that include about 300,000 items and 7 million responses. By providing this service online, we are able to collect a tremendous amount of data about how pairwise wiki surveys work in practice, and our steady stream of users provides a natural testbed for further methodological research.

The data collection process in a pairwise wiki survey is illustrated by a project conducted by the New York City Mayor's Office of Long-Term Planning and Sustainability in order to integrate residents' ideas into PlaNYC 2030, New York's citywide sustainability plan. The City has typically held public meetings and small focus groups to obtain feedback from the public. By using a pairwise wiki survey, the Mayor's Office sought to broaden the dialogue to include input from residents who do not traditionally attend public meetings. To begin the process, the Mayor's Office generated a list of 25 ideas based on their previous outreach (e.g., "Require all big buildings to make certain energy efficiency upgrades," "Teach kids about green issues as part of school curriculum").

Using these 25 ideas as "seeds," the Mayor's Office created a pairwise wiki survey with the question "Which do you think is a better idea for creating a greener, greater New York City?" Respondents were presented with a pair of ideas (e.g., "Open schoolyards across the city as public playgrounds" and "Increase targeted tree plantings in neighborhoods with high asthma rates"), and asked to choose between them (see Fig 2). After choosing, respondents were immediately presented with another randomly selected pair of ideas (the process for choosing the pairs is described in S1 Text). Respondents were able to continue contributing information about their preferences for as long as they wished by either voting or choosing "I can't decide." Crucially, at any point, respondents were able to contribute their own ideas, which—pending approval by the wiki survey creator—became part of the pool of ideas to be presented to others. Respondents were also able to view the popularity of the ideas at any time, making the process transparent. However, by decoupling the processes of voting and viewing the results—which occur on distinct screens (see Fig 2)—the site prevents a respondent from having immediate information about the opinions of others when she responds, which minimizes the risk of social influence and information cascades [43, 45–48].

The Mayor's Office launched its pairwise wiki survey in October 2010 in conjunction with a series of community meetings to obtain resident feedback. The effort was publicized at meetings in all five boroughs of the city and via social media. Over about four months, 1,436 respondents contributed 31,893 responses and 464 ideas to the pairwise wiki survey.

Data analysis

Given this data collection process, we analyze data from a pairwise wiki survey in two main steps (Fig 3). First, we use responses to estimate the opinion matrix Θ that includes an estimate of how much each respondent values each item. Next, we summarize the opinion matrix to produce a score for each item that estimates the probability that it will beat a randomly chosen item for a randomly chosen respondent. Because this analysis is modular, either step—estimation or summarization—could be improved independently.
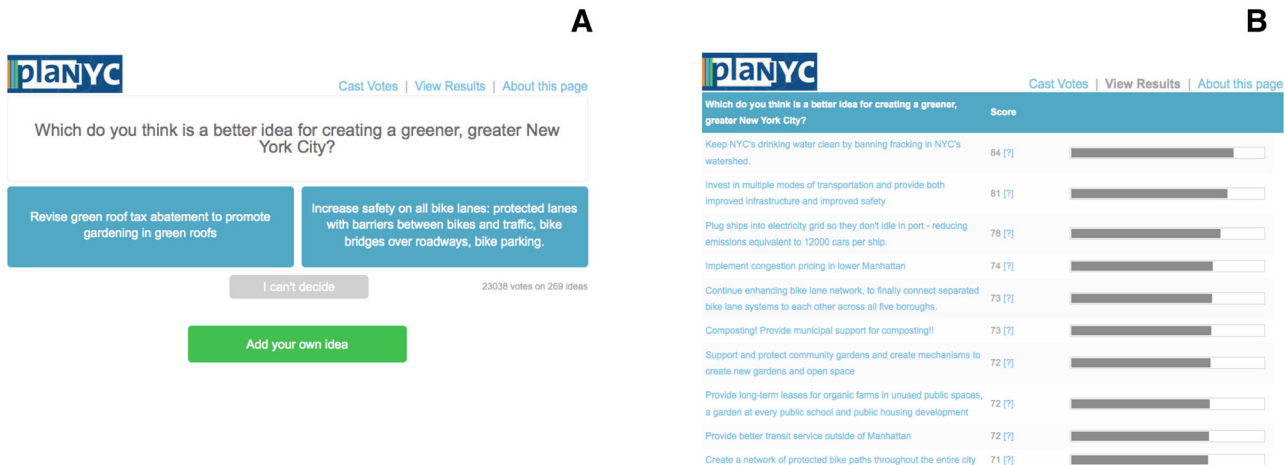
Fig 2. Response and results interfaces at www.allourideas.org. This example is from a pairwise wiki survey created by the New York City Mayor's Office to learn about residents' ideas about how to make New York "greener and greater."

Estimating the opinion matrix. The analysis begins with a set of pairwise comparison responses that are nested within respondents. For example, Fig 3 shows five hypothetical responses from two respondents. These responses are used to estimate the opinion matrix

$$
\Theta = \begin{bmatrix}
\theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,K} \\
\theta_{2,1} & \theta_{2,2} & \dots & \theta_{2,K} \\
\vdots & \vdots & \ddots & \vdots \\
\theta_{J,1} & \theta_{J,2} & \dots & \theta_{J,K}
\end{bmatrix}
$$

which has one row for each respondent and one column for each item, where $\theta_{j,k}$ is the amount that respondent j values item k (or more generally, the amount that respondent j believes item k answers the question being asked). In the New York City example described above, $\theta_{j,k}$ could be the amount that a specific respondent values the idea "Open schoolyards across the city as public playgrounds."
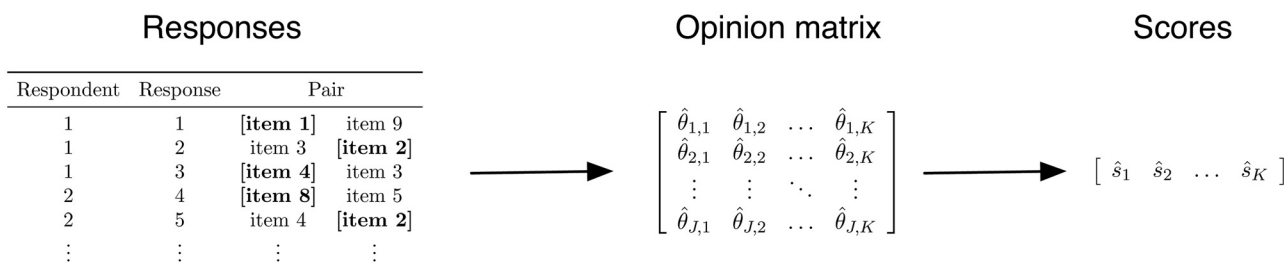

Fig 3. Summary of data analysis plan. We use responses to estimate the opinion matrix Θ and then we summarize the opinion matrix with the scores of each item.

Three features of the response data complicate the process of estimating the opinion matrix Θ. First, because the wiki survey is greedy, we have an unequal number of responses from each respondent. Second, because the wiki survey is collaborative, there are some items that can never be presented to some respondents. For example, if respondent j contributed an item, then none of the previous respondents could have seen that item. Collectively, the greediness and the collaborativeness mean that in practice we often have to estimate a respondent's value for an item that she has never encountered. The third problem is that responses are in the form of pairwise comparisons, which means that we can only observe a respondent's relative preference between two items, not her absolute feeling about either item.

In order to address these three challenges, we propose a statistical model that assumes that respondents' responses reflect their relative preferences between items (i.e., the Thurstone-Mosteller model [41, 49, 50]) and that the distribution of preferences across respondents for each item follows a normal distribution; see S2 Text for more information. Given these assumptions and weakly informative priors, we can perform Bayesian inference to estimate the $\theta_{j,\,k}$'s that are most consistent with the responses that we observe and the assumptions that we have made. One important feature of this modeling strategy is that for those who contribute many responses, we can better estimate their row in the opinion matrix, and for those who contribute fewer responses, we have to rely more on the pooling of information from other respondents (i.e., imputation). The specific functional forms that we assume (see S2 Text) result in the following posterior distribution, which resembles a hierarchical probit model:

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\mu} \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}) \quad &\propto \quad \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \times \prod_{j=1}^{J} \prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \\
&\times \prod_{k=1}^{K} N(\mu_k \mid \mu_{0[k]}, \tau_{0[k]}^2)
\end{aligned}
\tag{1}
$$

where X is an appropriately constructed design matrix, Y is an appropriately constructed outcome vector, $\boldsymbol{\mu} = \mu_1 \ldots \mu_K$ represents the mean appeal of each item, and $\boldsymbol{\mu_0} = \mu_{0[1]} \ldots \mu_{0[K]}$ and $\boldsymbol{\tau_0^2} = \tau_{0[1]}^2 \ldots \tau_{0[K]}^2$ are parameters to the priors for mean appeal of each item ($\boldsymbol{\mu}$).

This statistical model is just one of many possible approaches to estimating the opinion matrix from the response data, and we hope that future research will develop improved approaches. In S2 Text, we fully derive the model, discuss situations in which our modeling assumptions might not hold, and describe the Gibbs sampling approach that we use to make repeated draws from the posterior distribution. Computer code to make these draws was written in R [51] and utilized the following packages: plyr [52], multicore [53], bigmemory [54], truncnorm [55], testthat [56], Matrix [57], and matrixStats [58].

Summarizing opinion matrix. Once estimated, the opinion matrix Θ may include hundreds of thousands of parameters —there are often thousands of respondents and hundreds of items—that are measured on a non-intuitive scale. Therefore, the second step of our analysis is to summarize the opinion matrix Θ in order to make it more interpretable. The ideal summary of the opinion matrix will likely vary from setting to setting, but our preferred summary statistic is what we call the score of each item, $\widehat{s}_i$, which is the estimated chance that it will beat a randomly chosen item for a randomly chosen respondent. That is,

$$
\widehat{s}_i = \frac{\sum_{j=1}^{J} \sum_{k \neq i} \Phi(\hat{\theta}_{j,i} - \hat{\theta}_{j,k})}{J \times (K-1)} \times 100
\tag{2}
$$

The minimum score is 0 for an item that is always expected to lose, and the maximum score is 100 for an item that is always expected to win. For example, a score of 50 for the idea "Open schoolyards across the city as public playgrounds" means that we estimate it is equally likely to win or lose when compared to a randomly selected idea for a randomly selected respondent. To construct 95% posterior intervals around the estimated scores, we use the t posterior draws of the opinion matrix $(\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(t)})$ to calculate t posterior draws of $s(\widehat{s}^{(1)}, \widehat{s}^{(2)}, \ldots, \widehat{s}^{(t)})$. From these draws, we calculate the 95% posterior intervals around $\widehat{s}_i$ by findings values a and b such that $Pr(\widehat{s}_i > a) = 0.025$ and $Pr(\widehat{s}_i < b) = 0.025$ [59].

We chose to conduct a two-step analysis process—estimating and then summarizing the opinion matrix, Θ—rather than estimating the scores directly for three reasons. First, we believe that making the opinion matrix, Θ, an explicit target of inference underscores the possible heterogeneity of preferences among respondents. Second, by estimating the opinion matrix as an intermediate step, our approach can be extended to cases in which co-variates are added at the level of the respondent (e.g., gender, age, income, etc.) or at the level of the item (e.g., about the economy, about the environment, etc.). Finally, although we are currently most interested in the score as a summary statistic, there are many possible summaries of the opinion matrix that could be important, and by estimating Θ we enable future researchers to choose other summaries that may be important in their setting (e.g., which items cluster together such that people who value one item in the cluster tend to value other items in the cluster?). We return to some possible improvements, extensions, and generalizations in the Discussion.

Case studies

To show how pairwise wiki surveys operate in practice, in this section we describe two case studies in which the All Our Ideas platform was used for collecting and prioritizing community ideas for policymaking: New York City's PlaNYC 2030 and the Organisation for Economic Co-operation and Development (OECD)'s "Raise Your Hand" initiative. As described previously, the New York City Mayor's Office conducted a wiki survey in order to integrate residents' ideas into the 2011 update to the City's long-term sustainability plan. The wiki survey asked residents to contribute their ideas about how to create "a greener, greater New York City" and to vote on the ideas of others. The OECD's wiki survey was created in preparation for an Education Ministerial Meeting and an Education Policy Forum on "Investing in Skills for the 21st Century." The OECD sought to bring fresh ideas from the public to these events in a democratic, transparent, and bottom-up way by seeking input from education stakeholders located around the globe. To accomplish these goals, the OECD created a wiki survey to allow respondents to contribute and vote on ideas about "the most important action we need to take in education today."

We assisted the New York City Mayor's Office and the OECD in the process of setting up their wiki surveys, and spoke with officials of both institutions multiple times over the course of survey administration. We also conducted qualitative interviews with officials from both groups at the conclusion of survey data collection in order to better understand how the wiki surveys worked in practice, contextualize the results, and get a better sense of whether the use of a wiki survey enabled the groups to obtain information that might have been difficult to obtain via other data collection methods. Unfortunately, logistical considerations prevented either group from using a probabilistic sampling design. Therefore, we can only draw inferences about respondents, who should not be considered a random sample from some larger population. However, wiki surveys can be used in conjunction with probabilistic sampling designs, and we will return to the issue of sampling in the Discussion.
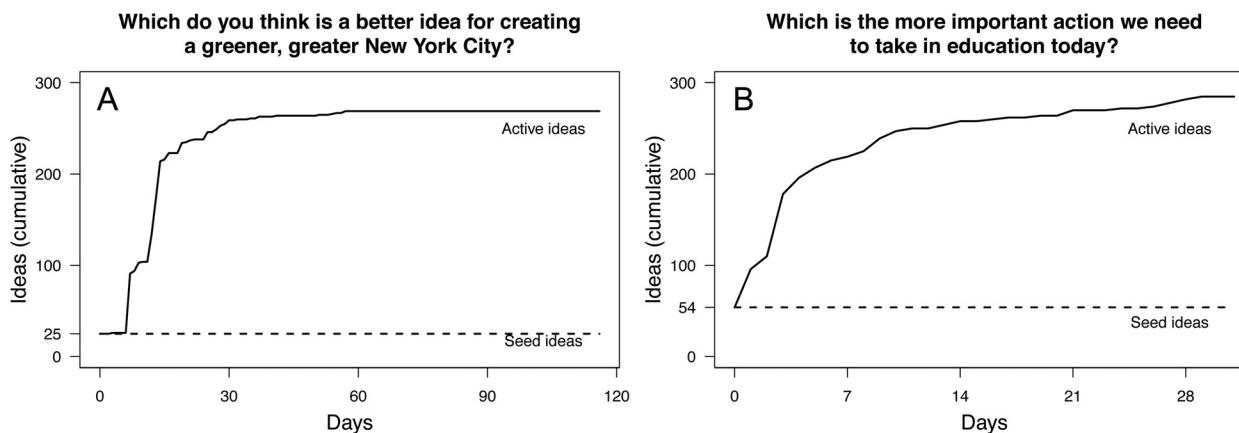
Fig 4. Cumulative number of activated ideas for PlaNYC [A] and OECD [B]. The PlaNYC wiki survey ran from October 7, 2010 to January 30, 2011. The OECD wiki survey ran from September 15, 2010 to October 15, 2010. In both cases the pool of ideas grew over time as respondents contributed to the wiki survey. PlaNYC had 25 seed ideas and 464 user-contributed ideas, 244 of which the Mayor's Office activated. The OECD had 60 seed ideas (6 of which it deactivated during the course of the survey), and 534 user-contributed ideas, 231 of which it activated. In both cases, ideas that were deemed inappropriate or duplicative were not activated.

Quantitative results

The pairwise wiki surveys conducted by the New York City Mayor's Office and the OECD had similar patterns of respondent participation. In the PlaNYC wiki survey, 1,436 respondents contributed 31,893 responses, and in the OECD wiki survey 1,668 respondents contributed 28,852 responses. Further, respondents contributed a substantial number of new ideas (464 for PlaNYC, and 534 for OECD). Of these contributed ideas, those that the wiki survey creators deemed inappropriate or duplicative were not activated. In the end, the number of ideas under consideration was dramatically expanded. For PlaNYC the number of active ideas in the wiki survey increased from 25 to 269, a 10-fold increase, and for the OECD from 60 to 285, a 5-fold increase (Fig 4).

Within each survey, the level of respondent contribution varied widely, in terms of both number of responses and number of ideas contributed, as we expected given the greedy nature of the wiki survey. In both cases, the distributions of both responses and contributed ideas contained "fat heads" and "long tails" (see Fig 5). If the wiki surveys captured only a fixed amount of information per respondent—as opposed to capturing all levels of effort—a significant amount of information would have been lost. For instance, if we only accepted the first 10 responses per respondent and discarded all respondents with fewer than 10 responses, approximately 75% of the responses in each survey would have been discarded. Further, if we were to limit the number of ideas contributed to one per respondent, as is typical in surveys with one and only one "other box," we would have excluded a significant number of new ideas: nearly half of the user-contributed ideas in the PlaNYC survey and about 40% in the OECD survey.

In both cases, many of the highest-scoring ideas were contributed by respondents. For PlaNYC, 8 of the top 10 ideas were contributed by users, as were 7 of the top 10 ideas for the OECD (Fig 6). These high-scoring user-contributed ideas highlight a strength of pairwise relative wiki surveys relative to both surveys and interviews. With a survey, it would have been difficult to learn about these new user-contributed ideas, and with an interview it would have been difficult to empirically assess the support that respondents have for them.
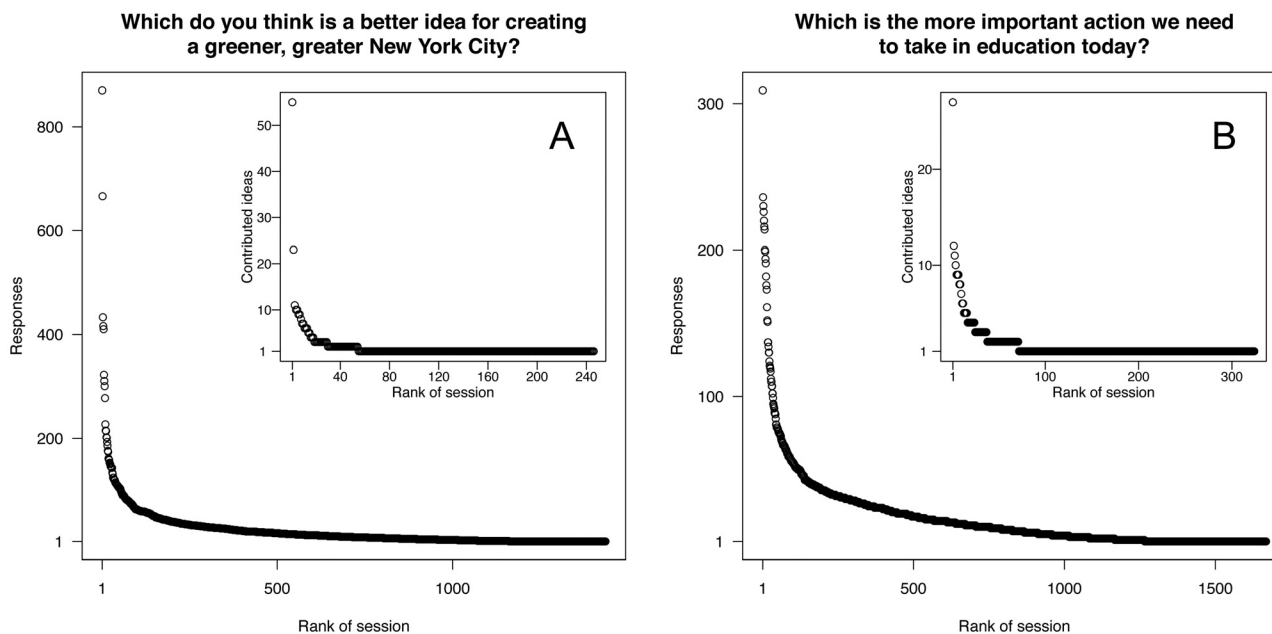
Which do you think is a better idea for creating a greener, greater New York City?

Which is the more important action we need to take in education today?

Fig 5. Distribution of contribution per respondent for PlaNYC [A] and OECD [B]. Both the number of responses per respondent and the number of ideas contributed per respondent show a "fat head" and a "long tail." Note that the scales on the figures are different.

Building on these specific results, we can begin to formulate a general model that describes the situations in which many of the top scoring items will be contributed by respondents. Three mathematical factors determine the extent to which an idea generation process will produce extreme outcomes (i.e., high scoring ideas): the number of ideas, the mean of ideas' scores, and the variance of ideas' scores [60]. In both of these case studies, there were many more user-contributed ideas than seed ideas, and they had higher variance in scores (Fig 7). These two features—volume and variance—ensured that many of the highest-scoring ideas were contributed by respondents, even though these ideas had a lower mean score than the seed ideas. Thus, in settings in which researchers seek to discover the highest-scoring ideas, the high variance and high volume of user-contributed ideas make them a likely source of these extreme outcomes.

Qualitative results

Because user-contributed ideas that score well are likely to be of interest—in fact, they highlight the value of the collaborativeness of wiki surveys—we sought to understand more about these items by conducting interviews with the creators of the PlaNYC and OECD wiki surveys. Based on these interviews, as well as interviews with six other wiki survey creators, we identified two general categories of high-scoring user-contributed ideas: *novel information*—that is, substantively new ideas that were not anticipated by the wiki survey creators—and *alternative framings*—that is, new and resonant ways of expressing existing ideas.

Some high-scoring user-contributed ideas contained information that was novel to the wiki survey creator. For example, in the PlaNYC context, the Mayor's Office reported that user-contributed ideas were sometimes able to bridge multiple policy arenas (or "silos") that might have been more difficult connections to make for office staff working within a specific arena. For instance, consider the high-scoring user-contributed idea "plug ships into electricity grid so they

Which do you think is a better idea for creating a greener, greater New York City?

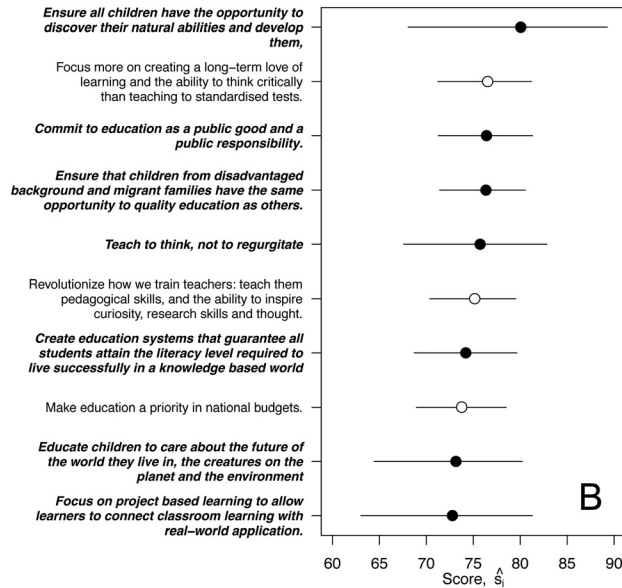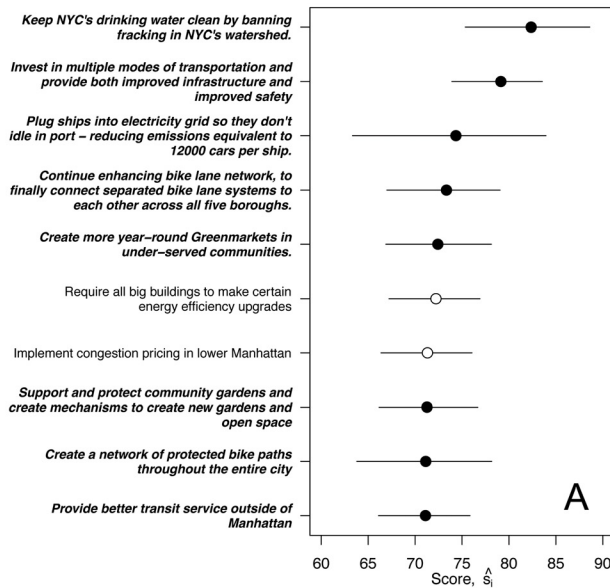Which is the more important action we need to take in education today?

Fig 6. Ten highest-scoring ideas for PlanNYC [A] and OECD [B]. Ideas that were contributed by respondents are printed in a bold/italic font and marked by closed circles; seed ideas are printed in a standard font and marked by open circles. In the case of PlaNYC, 8 of the 10 highest-scoring ideas were contributed by respondents. In the case of the OECD, 7 of the 10 highest-scoring ideas were contributed by respondents. Horizontal lines show 95% posterior intervals.

doi:10.1371/journal.pone.0123483.g006

don't idle in port—reducing emissions equivalent to 12000 cars per ship." The Mayor's Office suggested that staff may not have prioritized such an idea internally (it did not appear on the Mayor's Office's list of seed ideas), even though the idea's high score suggested public support for this policy goal: "[T]his relates to two areas. So plugging ships into electricity grid, so that's one, in terms of energy and sourcing energy. And it relates to freight. [Question: Okay, which are two separate silos?] Correct, so freight is something that we're looking closer at. . . . And
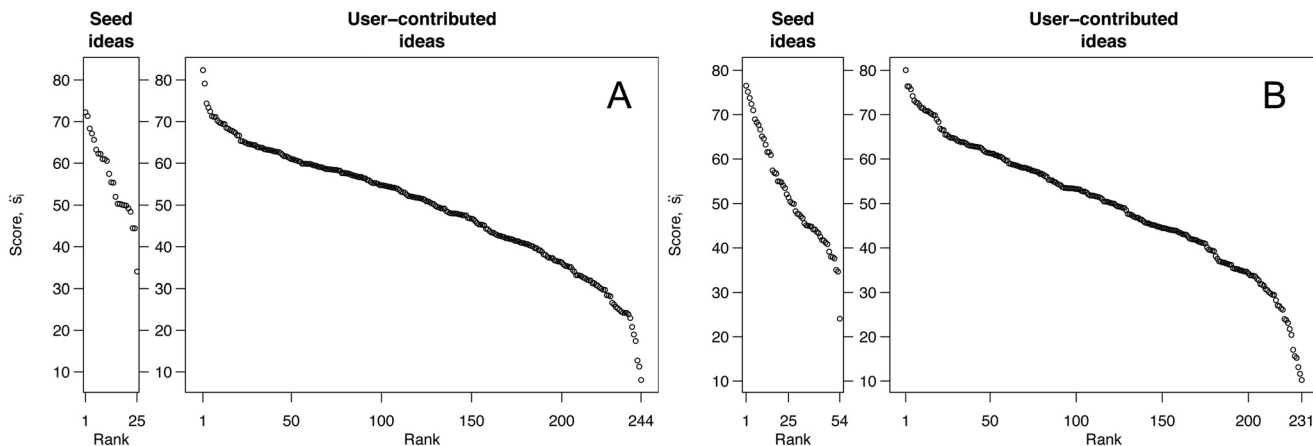

Fig 7. Distribution of scores of seed ideas and user-contributed ideas for PlaNYC [A] and OECD [B]. In both cases, some of the lowest-scoring ideas were user-contributed, but critically, some of the highest-scoring ideas were also user-contributed. In general, the large number of user-contributed ideas, combined with their high variance, means that they typically include some extremely popular ideas. Posterior intervals for each estimate are not shown.

doi:10.1371/journal.pone.0123483.g007

www.manaraa.com

emissions, reducing emissions, is something that's an overall goal of the plan. . . . So this has a lot of value to it for us to learn from" (interview with Ibrahim Abdul-Matin, New York City Mayor's Office, December 12, 2010).

Other user-contributed ideas suggested alternative framings for existing ideas. For instance, the creators of the OECD wiki survey noted that high-scoring, user-contributed ideas like "Teach to think, not to regurgitate" "wouldn't be formulated in such a way [by the OECD]. . . . [I]t's very un-OECD-speak, which we liked" (interview with Julie Harris, OECD, February 3, 2011). More generally, OECD staff noted that "what for me has been most interesting is that . . . those top priorities [are] very much couched in the language of principles[. . . .] It's sort of constitutional language" (interview with Joanne Caddy, OECD, February 15, 2011). PlaNYC's wiki survey creators also described the importance of user-contributed ideas being expressed in unexpected ways. The top-scoring idea in PlaNYC's wiki survey, contributed by a respondent, was "Keep NYC's drinking water clean by banning fracking in NYC's watershed"; Mayor's Office staff indicated that the office would have used more general language about protecting the watershed, rather than referencing fracking explicitly: "[W]e talk about it differently. We'll say, 'protect the watershed.' We don't say, 'protect the watershed from fracking'" (interview with Ibrahim Abdul-Matin, New York City Mayor's Office, December 12, 2010).

Taken together, these two case studies suggest that pairwise wiki surveys can provide information that is difficult, if not impossible, to gather from more traditional surveys or interviews. This unique information comes from high-scoring user-contributed ideas, and may involve both the content of the ideas and the language used to frame them.

Discussion

In this paper we propose a new class of data collection instruments called *wiki surveys*. By combining insights from traditional survey research and projects such as Wikipedia, we propose three general principles that all wiki surveys should satisfy: they should be greedy, collaborative, and adaptive. Designing an instrument that satisfies those three criteria introduces a number of challenges for data collection and data analysis, which we attempt to resolve in the form of a pairwise wiki survey. Through two case studies we show that pairwise wiki surveys can enable data collection that would be difficult with other methods. Moving beyond these proof-of-concept case studies to a fuller understanding of the strengths and weaknesses of pairwise wiki surveys, in particular, and wiki surveys, in general, will require substantial additional research.

One next step for improving our understanding of the measurement properties of pairwise wiki surveys would be additional studies to assess the consistency and validity of responses. Consistency could be assessed by measuring the extent to which respondents provide identical responses to the same pair and provide transitive responses to a series of pairs. Assessing validity would be more difficult, however, because wiki surveys tend to measure subjective states, such as attitudes, for which gold-standard measures rarely exist [61]. Despite the inherent difficulty of validating measures of subjective states, there are several approaches that could lead to increased confidence in the validity of pairwise wiki surveys [62]. First, studies could be done to assess discriminant validity by measuring the extent to which groups of respondents who are thought to have different preferences produce different wiki survey results. Second, construct validity could be assessed by measuring the extent to which responses for items that we believe to be similar are in fact similar. Third, studies could assess predictive validity by measuring the ability of results from pairwise wiki surveys to predict the future behavior of respondents. Finally, the results of pairwise wiki surveys could be compared to data collected through other quantitative and qualitative methodologies.

Another area for future research about pairwise wiki surveys is improving the statistical methods used to estimate the opinion matrix—either by choosing pairs more efficiently or developing more flexible statistical models. First, one could develop algorithms that would choose pairs so as to maximize the amount learned from each respondent. However, maximizing the amount of information per response [63–65] may not maximize the amount of information per respondent, which is determined by both the information per response and the number of responses provided by the respondent [66]. That is, an algorithm that chooses very informative pairs from a statistical perspective might not be effective if people do not enjoy responding to those kinds of pairs. Thus, algorithms could be developed to address both maximization of information per pair and to encourage participation by, for example, choosing pairs to which respondents enjoy responding. In addition to choosing pairs more efficiently, we believe that substantial progress can be made by developing more flexible and general statistical models for estimating the opinion matrix from a set of responses. For example, the statistical model we propose could be extended to include co-variates at the level of the respondent (e.g., age, gender, level of education, etc.) and at the level of the item (e.g., phrase structure, item topic, etc.). Another modeling improvement would involve creating more flexible assumptions about the distributions of opinions among respondents. These methodological improvements could be assessed by their robustness and their ability to improve the prediction of future responses (e.g., [67]).

Another important next step is to combine pairwise wiki surveys with probabilistic sampling methods, something that was logistically impossible in our case studies. If one thinks of survey research as a combination of sampling and interacting with respondents [68], then pairwise wiki surveys should be considered a new way of interacting with respondents, not a new way of sampling. However, pairwise wiki surveys can be naturally combined with a variety of different sampling designs. For example, researchers wishing to employ pairwise wiki surveys with a nationally representative sample can make use of commercially available online panels [69, 70]. Further, researchers wishing to study more specific groups—e.g., workers in a firm or residents in a city—could draw their own probability samples from administrative records.

Given the significant amount of work that remains to be done, we have taken a number of concrete steps to facilitate the future development of pairwise wiki surveys. First, we have made it easy for other researchers to create and host their own pairwise wiki surveys at www.allourideas.org. Further, the website enables researchers to download detailed data from their survey which can be analyzed in any way that researchers find appropriate. Finally, we have made all of the code that powers www.allourideas.org available open-source so that anyone can modify and improve it. We hope that these concrete steps will stimulate the development of pairwise wiki surveys. Further, we hope that other researchers will create different types of wiki surveys, particularly wiki surveys in which respondents themselves help to generate the questions [71, 72]. We expect that the development of wiki surveys will lead to new and powerful forms of open and quantifiable data collection.

Ethics Statement

All data collection and protection procedures were approved by the Institutional Review Board of Princeton University (protocol 4885).

Supporting Information

S1 Text. Additional description of the operation of the website.
(PDF)

S2 Text. Additional description of the data analysis procedure.
(PDF)

Acknowledgments

Author Contributions

References

1. Lazarsfeld PF. The Controversy Over Detailed Interviews—An Offer for Negotiation. Public Opinion Quarterly. 1944; 8(1):38–60. doi: 10.1086/265666

2. Converse JM. Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s. Public Opinion Quarterly. 1984; 48(1):267–282. doi: 10.1086/268825

3. Converse JM. Survey research in the United States: Roots and emergence 1890–1960. New Brunswick: Transaction Publishers; 2009.

4. Schuman H. Method and meaning in polls and surveys. Cambridge: Harvard University Press; 2008.

5. Schuman H, Presser S. The Open and Closed Question. American Sociological Review. 1979 Oct; 44 (5):692–712. doi: 10.2307/2094521

6. Schuman H, Scott J. Problems in the Use of Survey Questions to Measure Public Opinion. Science. 1987 May; 236(4804):957–959. doi: 10.1126/science.236.4804.957 PMID: 17812751

7. Presser S. Measurement Issues in the Study of Social Change. Social Forces. 1990 Mar; 68(3):856–868. doi: 10.2307/2579357

8. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 2014 Oct; 58(4):1064–1082. doi: 10.1111/ajps.12103

9. Krosnick JA. Survey Research. Annual Review of Psychology. 1999 Feb; 50(1):537–567. doi: 10.1146/annurev.psych.50.1.537 PMID: 15012463

10. Mitofsky WJ. Presidential Address: Methods and Standards: A Challenge for Change. Public Opinion Quarterly. 1989; 53(3):446–453. doi: 10.1093/poq/53.3.446

11. Dillman DA. Presidential Address: Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early Twenty-First Century. Public Opinion Quarterly. 2002 Oct; 66(3):473–494. doi: 10.1086/342184

12. Couper MP. Designing effective web surveys. Cambridge, UK: Cambridge University Press; 2008.

13. Couper MP, Miller PV. Web Survey Methods: Introduction. Public Opinion Quarterly. 2009 Jan; 72 (5):831–835. doi: 10.1093/poq/nfn066

14. Couper MP. The Future of Modes of Data Collection. Public Opinion Quarterly. 2011; 75(5):889–908. doi: 10.1093/poq/nfr046

15. Groves RM. Three Eras of Survey Research. Public Opinion Quarterly. 2011; 75(5):861–871. doi: 10.1093/poq/nfr057

16. Newport F. Presidential Address: Taking AAPOR's Mission To Heart. Public Opinion Quarterly. 2011; 75(3):593–604. doi: 10.1093/poq/nfr027

17. Benkler Y. The wealth of networks: How social production transforms markets and freedom. New Haven, CT: Yale University Press; 2006.

18. Howe J. Crowdsourcing: Why the power of the crowd is driving the future of business. New York: Three Rivers Press; 2009.

19. Noveck BS. Wiki government: How technology can make government better, democracy stronger, and citizens more powerful. Washington, D.C.: Brookings Institution Press; 2009.

20. Nielsen MA. Reinventing discovery: The new era of networked science. Princeton, N.J.: Princeton University Press; 2012.

21. Anderson C. The long tail: Why the future of business is selling less of more. New York, NY: Hyperion; 2006.

22. Wilkinson DM. Strong Regularities in Online Peer Production. In: Proceedings of the 9th ACM Conference on Electronic Commerce; 2008. p. 302–309.

23. Merton RK, Kendall PL. The Focused Interview. American Journal of Sociology. 1946 May; 51(6):541–557. doi: 10.1086/219886

24. Merton RK. The Focussed Interview and Focus Groups: Continuities and Discontinuities. Public Opinion Quarterly. 1987; 51(4):550–566. doi: 10.1086/269057

25. Presser S, Couper MP, Lessler JT, Martin E, Martin J, Rothgeb JM, et al. Methods for Testing and Evaluating Survey Questions. Public Opinion Quarterly. 2004 Apr; 68(1):109–130. doi: 10.1093/poq/nfh008

26. Balasubramanian SK, Kamakura WA. Measuring Consumer Attitudes Toward the Marketplace with Tailored Interviews. Journal of Marketing Research. 1989; 26(3):311–326. doi: 10.2307/3172903

27. Singh J, Howell RD, Rhoads GK. Adaptive Designs for Likert-type Data: An Approach for Implementing Marketing Surveys. Journal of Marketing Research. 1990; 27(3):304–321. doi: 10.2307/3172588

28. Groves RM, Heeringa SG. Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006; 169(3):439–457. doi: 10.1111/j.1467-985X.2006.00423.x

29. Toubia O, Flores L. Adaptive Idea Screening Using Consumers. Marketing Science. 2007; 26(3):342–360. doi: 10.1287/mksc.1070.0273

30. Smyth JD, Dillman DA, Christian LM, Mcbride M. Open-ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? Public Opinion Quarterly. 2009 Jun; 73(2):325–337. doi: 10.1093/poq/nfp029

31. Chen K, Chen H, Conway N, Hellerstein JM, Parikh TS. Usher: Improving Data Quality with Dynamic Forms. In: 2010 IEEE 26th International Conference on Data Engineering (ICDE); 2010. p. 321–332.

32. Dzyabura D, Hauser JR. Active Machine Learning for Consideration Heuristics. Marketing Science. 2011; 30(5):801–819. doi: 10.1287/mksc.1110.0660

33. Montgomery JM, Cutler J. Computerized Adaptive Testing for Public Opinion Surveys. Political Analysis. 2013 Apr; 21(2):172–192. doi: 10.1093/pan/mps060

34. Lewry F, Ryan T. Kittenwar: May the cutest kitten win! San Francisco: Chronicle Books; 2007.

35. Wu M. USG launches new web tool to gauge students' priorities. The Daily Princetonian. 2008.

36. Weinstein JR. Photocracy: Employing Pictoral Pair-Wise Comparison to Study National Identity in China, Japan, and the United States via the Web. Senior Thesis. Princeton University: Department of East Asian Studies; 2009.

37. Shah D. Solving Problems Using the Power of Many: Information Aggregation Websites, A Theoretical Framework and Efficacy Test. Senior Thesis. Princeton University: Department of Sociology; 2009.

38. Das Sarma A, Das Sarma A, Gollapudi S, Panigrahy R. Ranking mechanisms in twitter-like forums. In: Proceedings of the third ACM international conference on Web search and data mining. WSDM'10. New York, NY, USA: ACM; 2010. p. 21–30.

39. Luon Y, Aperjis C, Huberman BA. Rankr: A Mobile System for Crowdsourcing Opinions. In: Zhang JY, Wilkiewicz J, Nahapetian A, editors. Mobile Computing, Applications, and Services. No. 95 in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg; 2012. p. 20–31.

40. Salesses P, Schechtner K, Hidalgo CA. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. PLoS ONE. 2013 Jul; 8(7):e68400. doi: 10.1371/journal.pone.0068400 PMID: 23894301

41. Thurstone LL. The Method of Paired Comparisons for Social Values. Journal of Abnormal and Social Psychology. 1927; 21(4):384–400. doi: 10.1037/h0065439

42. Hacker S, von Ahn L. Matchin: Eliciting User Preferences with an Online Game. Proceedings of the 27th international conference on Human factors in computing systems. 2009;p. 1207–1216.

43. Salganik MJ, Watts DJ. {Web-based} Experiments for the Study of Collective Social Dynamics in Cultural Markets. Topics in Cognitive Science. 2009 Jul; 1(3):439–468. doi: 10.1111/j.1756-8765.2009.01030.x PMID: 25164996

44. Goel S, Mason W, Watts DJ. Real and perceived attitude agreement in social networks. Journal of Personality and Social Psychology. 2010; 99(4):611–621. doi: 10.1037/a0020697 PMID: 20731500

45. Salganik MJ, Dodds PS, Watts DJ. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science. 2006 Feb; 311(5762):854–856. doi: 10.1126/science.1121066 PMID: 16469928

46. Zhu H, Huberman B, Luon Y. To switch or not to switch: understanding social influence in online choices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'12. New York, NY, USA: ACM; 2012. p. 2257–2266.

47. Muchnik L, Aral S, Taylor SJ. Social Influence Bias: A Randomized Experiment. Science. 2013 Aug; 341(6146):647–651. doi: 10.1126/science.1240466 PMID: 23929980

48. van de Rijt A, Kang SM, Restivo M, Patil A. Field experiments of success-breeds-success dynamics. Proceedings of the National Academy of Sciences. 2014 May; 111(19):6934–6939. doi: 10.1073/pnas.1316836111

49. Mosteller F. Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations. Psychometrika. 1951 Mar; 16:3–9. doi: 10.1007/BF02289116

50. Stern H. A Continuum of Paired Comparisons Models. Biometrika. 1990 Jun; 77(2):265–273. doi: 10.1093/biomet/77.2.265

51. R Core Team. R: A Language and Environment for Statistical Computing; 2014. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/

52. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software. 2011; 40(1):1–29.

53. Urbanek S. multicore: Parallel Processing of R Code on Machines with Multiple Cores or CPUs; 2011. R package version 0.1–5.

54. Kane MJ, Emerson JW. bigmemory: Manage Massive Matrices With Shared Memory and Memory-Mapped Files; 2011. R package version 4.2.11.

55. Trautmann H, Steuer D, Mersmann O, Bornkamp B. truncnorm: Truncated Normal Distribution; 2011. R package, version 1.0–5.

56. Wickham H. testthat: Get Started with Testing. The R Journal. 2011; 3(1):5–10.

57. Bates D, Maechler M. Matrix: Sparse and Dense Matrix Classes and Methods; 2011. R package version 0.999375–50.

58. Bengtsson H. matrixStats: Methods that apply to rows and columns of a matrix.; 2013. R package version 0.8.14.

59. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2003.

60. Girotra K, Terwiesch C, Ulrich KT. Idea Generation and the Quality of the Best Idea. Management Science. 2010 Apr; 56(4):591–605. doi: 10.1287/mnsc.1090.1144

61. Turner CF, Martin E. Surveying subjective phenomena. New York: Russell Sage Foundation; 1984.

62. Fowler FJ. Improving survey questions: design and evaluation. Thousand Oaks: Sage; 1995.

63. Lindley DV. On a Measure of the Information Provided by an Experiment. The Annals of Mathematical Statistics. 1956 Dec; 27(4):986–1005. doi: 10.1214/aoms/1177728069

64. Glickman ME, Jensen ST. Adaptive paired comparison design. Journal of Statistical Planning and Inference. 2005 Jan; 127(1–2):279–293. doi: 10.1016/j.jspi.2003.09.022

65. Pfeiffer T, Gao XA, Chen Y, Mao A, Rand DG. Adaptive Polling for Information Aggregation. In: Twenty-Sixth AAAI Conference on Artificial Intelligence; 2012.

66. von Ahn L, Dabbish L. Designing Games with a Purpose. Communications of the ACM. 2008; 51 (8):58–67. doi: 10.1145/1378704.1378719

67. Mao A, Soufiani HA, Chen Y, Parkes DC. Capturing Cognitive Aspects of Human Judgment; 2013. 1311.0251. Available from: http://arxiv.org/abs/1311.0251.

68. Conrad FG, Schober MF. Envisioning the survey interview of the future. Hoboken, N.J.: Wiley-Interscience; 2008.

69. Baker R, Blumberg SJ, Brick JM, Couper MP, Courtright M, Dennis JM, et al. Research Synthesis: AAPOR Report on Online Panels. Public Opinion Quarterly. 2010 Dec; 74(4):711–781. doi: 10.1093/poq/nfq048

70. Brick JM. The Future of Survey Sampling. Public Opinion Quarterly. 2011 Dec; 75(5):872–888. doi: 10.1093/poq/nfr045

71. Sullivan JL, Piereson J, Marcus GE. An Alternative Conceptualization of Political Tolerance: Illusory Increases 1950s-1970s. American Political Science Review. 1979; 73(3):781–794. doi: 10.2307/1955404

72. Gal D, Rucker DD. Answering the Unasked Question: Response Substitution in Consumer Surveys. Journal of Marketing Research. 2011 Feb; 48:185–195. doi: 10.1509/jmkr.48.1.185