Wiki surveys: Open and quantifiable social data collection*

Matthew J. Salganik† and Karen E.C. Levy‡

Abstract

Research about attitudes and opinions is central to social science and relies on two common methodological approaches: surveys and interviews. While surveys enable the quantification of large amounts of information quickly and at a reasonable cost, they are routinely criticized for being "top-down" and rigid. In contrast, interviews allow unanticipated information to "bubble up" directly from respondents, but are slow, expensive, and difficult to quantify. Advances in computing technology now enable a hybrid approach that combines the quantifiability of a survey and the openness of an interview; we call this new class of data collection tools *wiki surveys*. Drawing on principles underlying successful information aggregation projects, such as Wikipedia, we propose three general criteria that wiki surveys should satisfy: they should be greedy, collaborative, and adaptive. We then present results from `www.allourideas.org`, a free and open-source website we created that enables groups all over the world to deploy wiki surveys. To date, about 1,500 wiki surveys have been created, and they have collected over 60,000 ideas and 2.5 million votes. We describe the methodological challenges involved in collecting and analyzing this type of data and present case studies of wiki surveys created by the New York City Mayor's Office and the Organisation for Economic Co-operation and Development (OECD). We conclude with a discussion of limitations, many of which may be overcome with additional research.

*This research was supported by grants from Google (Faculty Research Award, the People and Innovation Lab, and Summer of Code 2010); Princeton University Center for Information Technology Policy; Princeton University Committee on Research in the Humanities and Social Sciences; the National Science Foundation [grant number CNS-0905086]; and the National Institutes of Health [grant number R24 HD047879]. We thank Peter Lubell-Doughtie, Adam Sanders, Pius Uzamere, Dhruv Kapadia, Chap Ambrose, Calvin Lee, Dmitri Garbuzov, Brian Tubergen, Peter Green, and Luke Baker for outstanding web development. Also, we thank Nadia Heninger, Bill Zeller, Bambi Tsui, Dhwani Shah, Gary Fine, Mark Newman, Dennis Feehan, Sophia Li, Lauren Senesac, Devah Pager, Paul DiMaggio, Adam Slez, and Scott Lynch for valuable suggestions, and Josh Weinstein for his critical role in the genesis of this project. Finally, we thank Ibrahim Abdul-Matin and colleagues at the New York City Mayor's Office and Joanne Caddy, Julie Harris, and Cassandra Davis at the Organisation for Economic Co-operation and Development. This work represent the views of the authors and not the users or funders of `www.allourideas.org`.

1 Introduction

Research about attitudes and opinions is central to social science and typically relies on two common methodological approaches: surveys and interviews. While surveys allow researchers to quantify large amounts of information quickly and at a reasonable cost, they are routinely criticized for being "top-down" and rigid; that is, the survey questions and possible responses are formulated *before* data collection begins, meaning that surveys generally are not open to novel or unexpected information from respondents. In contrast, interviews allow new information to "bubble up" directly from respondents, but are slow, expensive, and difficult to quantify.

This tension between openness and quantifiability, which underlies many disputes between quantitative and qualitative researchers, has a long history in survey research. During World War II, some of the nation's leading social scientists worked at the United States Office of War Information and were charged with measuring citizens' attitudes toward current events. Within the Office, fierce methodological debates arose about the best way to conduct this research (Converse, 1984, 2009). One group of researchers, led by Rensis Likert, preferred an open interview technique, to allow for nuanced assessment of attitudes; the other, influenced by commercial pollsters like Elmo Roper and George Gallup, favored closed questions that were easily administered and quickly quantified. Paul Lazarsfeld, perhaps the most eminent methodologist of the day, was called in to adjudicate the conflict, and wrote a review of the controversy that came down in favor of closed questions (Lazarsfeld, 1944). While each method of research had its proper role, Lazarsfeld wrote, closed questions were nearly as effective as open questions for most purposes, and were far more manageable to administer. It is largely on this practical basis that closed questions have come to dominate survey research, despite the strength of open-ended questions for collecting some types of information. Yet Lazarsfeld himself wondered: "Is there not some way to use all the good ideas which the proponents of the [open interview] technique have and still to develop methods that are more objective, more manageable on a mass basis . . . ?" (Lazarsfeld, 1944, p. 50).

We believe that such an approach is now possible due to the developments in computation and human connectivity brought about by the World Wide Web. Drawing from principles underlying successful information aggregation projects such as Wikipedia, we propose a new hybrid class of data collection tools—which we call *wiki surveys*—that integrates the quantifiability of surveys and the openness of interviews. After providing some background in Section 2, in Section 3 we describe three general properties that characterize wiki surveys: they should be greedy, collaborative, and adaptive. In Section 4 we discuss one specific type of wiki survey—a pairwise wiki survey—and present results from `www.allourideas.org`, a free and open-source website we created that allows groups all over the world to create and use pairwise wiki surveys. In Section 5, we propose a statistical model for estimating public opinion from pairwise wiki survey data, and in Section 6 we present two case studies of institutions that have used pairwise wiki surveys: the New York City Mayor's Office and the Organisation for Economic Co-operation and Development (OECD). The paper concludes with a discussion of the limitations of this work, many of which may be overcome with additional research.

2 Background

Closed questions, which have come to dominate survey research since World War II (Smith, 1987), provide a finite, predetermined set of answers among which respondents are asked to choose. The primary practical advantage of closed questions is that responses can be handled with relative ease: answers can be easily assigned values, fed into statistical software, and employed in quantitative analysis. These processes are relatively straightforward, fast, and inexpensive, making closed questions an efficient choice for large-scale social science surveys.

In contrast, responses to open-ended questions are more complicated for researchers to reliably code and quantify. In some cases, however, open methods may provide insights that closed methods cannot because they are receptive to new information that was unanticipated by the researcher (Schuman and Presser, 1979; Schuman and Scott, 1987; Presser, 1990; Schuman, 2008). For example, Schuman and Presser (1979) conducted a split-ballot test of an open and closed form of a question about what people value in jobs. 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. Because respondents have a strong tendency to confine their responses to the answer choices offered (Krosnick, 1999; Schuman, 2008), researchers who construct all the possible answer choices necessarily constrain what can be learned.[1] This is unfortunate because unanticipated information is often the most valuable for research.

The use of closed versus open approaches, then, represents a tradeoff: open approaches can potentially yield richer information, but this comes at the cost of being more onerous. However, new technologies present opportunities for hybrid survey methodologies that integrate the positive aspects of both closed and open approaches. The World Wide Web is an ideal platform for new and innovative forms of social data collection, and important work has been done to develop web-based surveys (Couper, 2008; Couper and Miller, 2009). Much of this work has been translational (Skitka and Sargis, 2006) in the sense that it attempts to move traditional approaches online (e.g., Couper et al. (2004); Christian et al. (2007); Smyth et al. (2009); Holland and Christian (2009); Yan et al. (2011)) and enrich traditional approaches using the possibilities available on the Web (e.g., Callegaro and DiSogra (2008); Shih and Fan (2008); Dillman et al. (2009); Farrell and Petersen (2010)). However, the Web also allows for new forms of data collection that have no obvious offline analog.

3 Wiki surveys

If one views a survey as a tool for information aggregation, then insights about how to conduct web-based surveys might be gleaned from successful web-based information aggregation projects, of which Wikipedia is an exemplar. Because our approach attempts to combine insights from projects such as Wikipedia with insights from survey research, we call our new instrument a *wiki survey*.

Benkler (2006) notes that successful information aggregation systems are typically composed of granular, modular tasks. That is, in successful systems, large problems can be broken down into smaller "chunks," which require low individual investment of time and effort (granularity), and these "chunks" can be independently completed by many individuals before being flexibly aggregated into a larger whole (modularity). If this insight is applied to survey research, we would want to ensure that each unit of information collected requires a very small investment of energy by a respondent, and that these "chunks" of information could be flexibly combined to produce measures of public opinion. Building on this insight, we believe that a wiki survey should have three primary characteristics: it should be *greedy*, *collaborative*, and *adaptive*.

[1] It is the case that researchers are often advised to use open methods to construct the set of responses for closed questions before primary data collection begins (Krause, 2002). However, in practice this can be quite difficult (Schuman and Presser, 1979), and, once determined, these response sets are typically not amenable to new information that may arise. Thus, any omissions that occur while designing the response set may limit what can be learned.

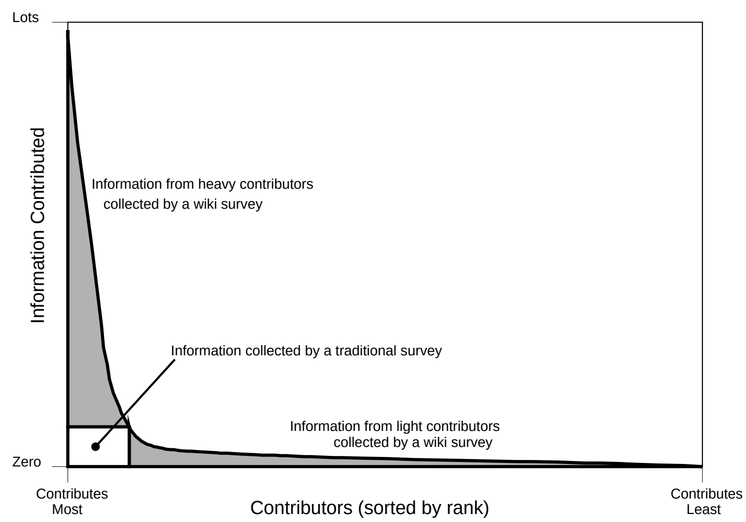


Figure 1: Schematic of rank order plot of contributions to successful information aggregation systems on the Web. These systems can handle both heavy contributors ("the fat head"), shown on the left side of the plot, and light contributors ("the long tail"), shows on the right side of the plot. Traditional survey methods utilize information from neither the "fat head" nor the "long tail" and thus leave huge amounts of information uncollected.

3.1 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 so. This contrasts sharply with successful information aggregation projects on the Internet, which collect as much or as little information as each respondent 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" (Anderson, 2006; Wilkinson, 2008) (Fig. 1). For example, the number of edits to Wikipedia per editor roughly follows a power-law distribution with an exponent 2 (Wilkinson, 2008). 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. Of course, such differential participation introduces challenges for analysis, but surely it is preferable to collect this information than not to do so.

3.2 Collaborativeness

In traditional surveys, the questions and answer choices are typically written by researchers and not 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 (Merton and Kendall, 1946; Merton, 1987). Thus, just as a community collaboratively writes and edits Wikipedia, the content of a wiki survey should be partially created by its users. This approach to collaborative survey construction resembles some forms of survey pre-testing (Presser et al., 2004). However, rather than thinking of pre-testing as a phase distinct from the "real" data collection, in wiki surveys the collaboration process continues throughout data collection.

3.3 Adaptivity

Traditional surveys are static: the question 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 for estimating the parameters of interest, given what is already known.[2] In other words, while collaborativeness involves being open to new information, adaptivity involves using the information that you already have more efficiently. In the context of wiki surveys, adaptivity is particularly important given that respondents will 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, based on experiences in other areas of research (e.g., standardized testing of students (van der Linden and Glas, 2010)), we believe that gains in efficiency from adaptivity can more than offset the cost of added complexity.

4 Pairwise Wiki Surveys

Our first attempt to operationalize these three principles of wiki surveys resulted in what we call a *pairwise wiki survey.* It 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 (Thurstone, 1927), is ideal for wiki surveys because it is amenable to the three criteria described above. Pairwise comparison can be *greedy* because one can easily present as many (or as few) prompts as each participant is willing to answer. New responses contributed by participants can be easily integrated into the choice sets of future respondents, enabling the instrument to be *collaborative.* Finally, it can be *adaptive* because researchers select the prompts to be evaluated, so we can select the prompts that are most informative given what has already been learned. These properties exist because pairwise comparison is both granular and modular: the unit of contribution is small and can be readily aggregated.

In addition, pairwise comparison has several practical advantages. First, pairwise comparison makes manipulation, or "gaming," of results difficult because respondents cannot choose which prompts they will see; instead, this choice is made by our system. Thus, when there is a large number of possible items, a respondent would have to vote many times in order to even be presented with the item that she wishes to "vote up" (or "vote down") (Hacker and von Ahn, 2009). Second, pairwise comparison requires respondents to prioritize possible choices, preventing them

[2]For related work on adaptive approaches to other types of surveys, see Balasubramanian and Kamakura (1989); Singh et al. (1990); Groves and Heeringa (2006); Toubia and Flores (2007); Smyth et al. (2009); Chen et al. (2010); Dzyabura and Hauser (2011).

from "liking everything"—that is, because the respondent must select one of two discrete answer choices for each prompt, 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, voting on a series of pairwise comparisons is reasonably enjoyable, a common characteristic of many successful web-based social research projects (Salganik and Watts, 2009). Perhaps because of these characteristics, pairwise comparison tools have become popular on the Web (e.g., Lewry and Ryan (2007); Weinstein (2009)).

4.1 All Our Ideas

Because no system for deploying pairwise wiki surveys existed, we created the website All Our Ideas (`www.allourideas.org`), described to visitors as "a tool for collecting and prioritizing ideas in a democratic, transparent, and efficient manner." Any visitor to the site can create an "idea marketplace" (i.e., a pairwise wiki survey) though which members of a group can vote on the ideas of others, via a series of comparisons, as well as submit their own ideas to the pool of available choices.[3] This allows groups to both collect and prioritize information in the same process. By providing this service freely on the Web, 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.

All Our Ideas is illustrated by a project that we conducted with 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 All Our Ideas, 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 wiki survey with the question "Which do you think is a better idea for creating a greener, greater New York City?" Participants in the wiki survey 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 Figure 2). After making this vote, participants were immediately presented with another randomly selected pair of ideas (the process for choosing the prompts is described in more detail in Appendix A). Voting continued for as long as the participant wished to contribute information about her preferences by voting or choosing "I can't decide." Crucially, at any point, participants were able to upload their own ideas, which then (pending approval by the wiki survey creator) became part of the pool of ideas to be voted on by others.

The Mayor's Office launched its 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 and via online and social media (mention on local and national blogs, on Facebook and Twitter, etc.). Over about four months, the wiki survey collected 28,829 votes and 464 user-submitted ideas; participation peaked in the days following media mentions and email follow-ups after public meetings. As these pairwise votes accumulated, the aggregate preferences of participants became evident. These results were visible to any visitor to the site in real time via a "view results" tab, making the process transparent (Fig. 2).[4] More about the statistical methods

[3] A related website, "Which Do You Want More?", was used by the Princeton University Undergraduate Student Government in 2009. For more on that project, see Weinstein (2009) and Shah (2009).

[4] By decoupling the processes of voting and viewing the results—which occur on distinct screens (see Fig. 2)—

(a) Voting Screen (b) Results Screen

Figure 2: Voting and results interfaces at `www.allourideas.org`. This example is from a 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."

underlying the website is presented in Section 5 and Appendix A, and more about the results of the New York City wiki survey is presented in Section 6.

This example illustrates the workings of just one wiki survey. Since `www.allourideas.org` launched in February 2010, about 1,500 wiki surveys have been created, which have collected 60,000 ideas and 2.5 million votes. Some of the most active wiki surveys were created by groups such as *The Washington Post*, Catholic Relief Services, Humanity in Action, the Craigslist Foundation, a Congressional candidate, the City of Calgary, the New York City Department of Parks and Recreation, the Occupy Wall Street movement, and student governments at Princeton and Columbia.

The website implementation requires algorithms for 1) choosing which prompts to show voters and 2) aggregating the votes to estimate a measure of public opinion. When constructing such algorithms for the website, we were limited to approaches that were simple enough computationally that they could be implemented in real time. The procedures that we employ on the website are described in Appendix A and are reasonable heuristic approaches that are consistent with the three principles of wiki surveys. Next, however, we turn our attention to a statistical approach for estimating public opinion from the votes that is currently too computationally intensive to be done in real time. We are confident, however, that the procedures running on the website will be improved in the future, and we will return to this issue in Section 7.

the site prevents a visitor from having immediate information on the preferences of others when she votes. This decoupling prevents "groupthink" and information cascades, whereby popularity can become a poor indicator of the underlying quality of the idea (Salganik et al., 2006; Salganik and Watts, 2009).

5 Data analysis

The main goal of the pairwise wiki survey is to estimate "public opinion" from the set of pairwise votes. More specifically, we conceptualize public opinion as a matrix $\mathbf{\Theta}$,

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

where $\theta_{j,k}$ is the amount that the person in session j likes item k. In the New York City example described above, this could be the amount that a specific voter likes the item "Open schoolyards across the city as public playgrounds."

If we can estimate how much each voter likes each item, then we can aggregate those estimates to describe the public opinion of the group. Unfortunately, because the responses are in the form of pairwise comparisons, they are not directly informative about how much each voter likes each item. That is, from each vote we can only observe the relative preference for two items, not the absolute preference for either. Therefore, we assume a model of the data-generating process and then use Bayesian inference to estimate the preferences most consistent with the observed data.

A full treatment of model development and the parameter estimation algorithm is provided in Appendix B, and a summary is provided here. First, for the process of voting, we assume that

$$Pr(\text{item } a \text{ beats item } b \text{ in session } j) = \Phi(\theta_{j,a} - \theta_{j,b}) \tag{2}$$

where Φ is a cumulative standard normal (Thurstone, 1927; David, 1988; Bockenholt, 2007). In other words, the greater the difference between $\theta_{j,a}$ and $\theta_{j,b}$ the more likely the person in user-session j will choose item a over item b.[5]

Given this model for the voting process, the likelihood can be written to resemble a standard probit model given an approximately constructed design matrix $\boldsymbol{X}$, outcome vector $\boldsymbol{Y}$, and parameter vector $\boldsymbol{\theta}$ (a full derivation is provided in Appendix B),

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T\boldsymbol{\theta})^{y_i}(1 - \Phi(\boldsymbol{x}_i^T\boldsymbol{\theta}))^{1-y_i} \tag{3}$$

Next, we add hierarchical structure to the model by assuming that the amount that each user-session likes a specific item (k) is drawn from a normal distribution with mean (μ_k) and standard deviation (σ). This results in the likelihood,

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\mu}, \sigma) = \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T\boldsymbol{\theta})^{y_i}(1 - \Phi(\boldsymbol{x}_i^T\boldsymbol{\theta}))^{1-y_i} \times \prod_{j=1}^{J}\prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \tag{4}$$

where $\boldsymbol{\mu} = \mu_1 \ldots \mu_K$ and σ is set to be 1.

[5]We use the term "user-session" rather than "respondent" to emphasize the fact that each user-session does not necessarily represent a unique respondent. For example, a respondent who participates at work and then later participates at home would be counted as two user-sessions. More specifically, a user-session is created when a browser that is not currently in a session visits the site. If there are 10 minutes of inactivity on the site, the current session is terminated; future activity on the site would result in a new session being created. The sessions are tracked with browser cookies, so a user could delete his or her cookies or open a different browser to create a new session.

Finally, we add conjugate priors to yield the following posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\mu} \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}) \quad \propto \quad & \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \times \prod_{j=1}^{J} \prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \\
& \times \prod_{k=1}^{K} N(\mu_k \mid \mu_{0[k]}, \tau_{0[k]}^2)
\end{aligned} \tag{5}
$$

where $\boldsymbol{\mu_0} = \mu_{0[1]} \ldots \mu_{0[K]}$ 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 ($\boldsymbol{\mu}$). In Appendix B, we describe the Gibbs sampling approach that we used to make repeated draws from the posterior distribution. These draws allow us to estimate the values of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ that are most consistent with our data, given our model.

Because the parameter vector $\boldsymbol{\theta}$ is just a vector representation of the public opinion matrix $\boldsymbol{\Theta}$, it would seem that once we can estimate $\boldsymbol{\theta}$ our task would be complete. Unfortunately, however, the estimated public opinion matrix $\widehat{\boldsymbol{\Theta}}$ is very difficult to interpret because it is quite large, often hundreds of thousands of parameters, and because $\widehat{\theta_{i,j}}$ is not measured on a natural scale. Therefore, we map the estimated public opinion matrix to a vector which stores the estimated *score* of each item. The score of an item is the estimated chance that it will beat a randomly chosen item for a randomly chosen user-session. That is,

$$
\widehat{s}_i = \frac{\sum_{j=1}^{J} \sum_{k \neq i} \Phi(\widehat{\theta_{j,i}} - \widehat{\theta_{j,k}})}{J \times (K-1)} \times 100 \tag{6}
$$

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 user-session. To construct 95% posterior intervals around the estimated scores, we use the t posterior draws of the public opinion matrix $(\boldsymbol{\Theta^{(1)}}, \boldsymbol{\Theta^{(2)}}, \ldots, \boldsymbol{\Theta^{(t)}})$ to calculate t posterior draws of $\boldsymbol{s}$ $(\widehat{\boldsymbol{s}}^{(1)}, \widehat{\boldsymbol{s}}^{(2)}, \ldots, \widehat{\boldsymbol{s}}^{(t)})$. From these draws, we calculate the 95% posterior intervals around $\widehat{s}_i$ in the usual way, by findings values a and b such that $Pr(\widehat{s}_i > a) = 0.025$ and $Pr(\widehat{s}_i < b) = 0.025$ (Gelman et al., 2003).

We chose to estimate the public opinion matrix, $\boldsymbol{\Theta}$, and then calculate the scores, $\widehat{\boldsymbol{s}}$, rather than estimate the scores directly for three reasons. First, this approach naturally handles the unequal amount of information that we have for each respondent due to the greediness of the pairwise wiki survey: for those who cast many votes, we can better estimate their row in the public opinion matrix, and for those who cast fewer votes, we have to rely more on the pooling of information from other respondents. Second, this approach can be generalized to cases where 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.). Third, explicit estimation of the full public opinion matrix also lends itself to calculation of other interesting aspects of public opinion (e.g., Which items cluster together such that people who like one item in the cluster tend to like other items in the cluster?). We return to some possible extensions and generalizations in Section 7. Finally, we note that this modeling approach assumes that the sample design is ignorable, but it can be extended to produce both estimates and confidence intervals under more general, non-ignorable sampling designs (e.g., stratified sampling, cluster sampling), assuming that the sample design is known (Gelman et al., 2003, Ch. 7).

6 Case studies

To understand 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. The All Our Ideas platform is well-suited to open government goals because of its openness, simplicity, and transparency: any site visitor can vote, contribute ideas, and view all results. Further, ideas can be evaluated objectively without concern for who suggested them.

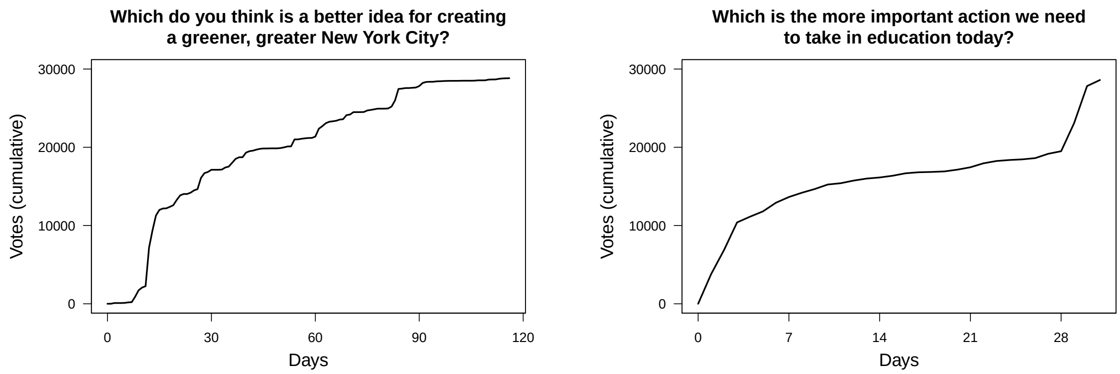
As described previously, the New York City Mayor's Office conducted a wiki survey in order to integrate residents' ideas into the 2011 update to the City's long-term sustainability plan. The wiki survey asked residents to upload their ideas about how to create "a greener, greater New York City" and to 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 submit 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 methods of data collection. Unfortunately, logistical considerations prevented either group used a probabilistic sampling design. Therefore, we can only draw inferences regarding the people who visited the website, and these people should not be considered a random sample from some larger population. Further, given the data collection in these two case studies, we cannot calculate a response rate. However, we emphasize that although it was not done in these two cases, wiki surveys can naturally be used in conjunction with probabilistic sampling designs; we will return to this issue in Section 7.

6.1 Results

The pairwise wiki surveys of the New York City Mayor's Office and the OECD were similar in scale, both in terms of number of user-submitted ideas and number of votes. Over about four months (Oct. 7, 2010 to Jan. 30, 2011), New York's PlaNYC wiki survey collected 28,829 votes and 464 user-submitted ideas, 244 of which the Mayor's Office activated; ideas that were deemed inappropriate or duplicative by the Mayor's office were not activated (see Figs. 3(a) and 4(a)). Similarly, over about a month (Sep. 15, 2010 to Oct. 15, 2010), the OECD's wiki survey collected 28,599 votes and 533 user-submitted ideas, 231 of which were activated by the OECD (see Figs. 3(b) and 4(b)). One difference between the two was the OECD wiki survey was global: votes were cast from more than ninety different countries and ideas were submitted from more than fifty countries (Fig. 5).

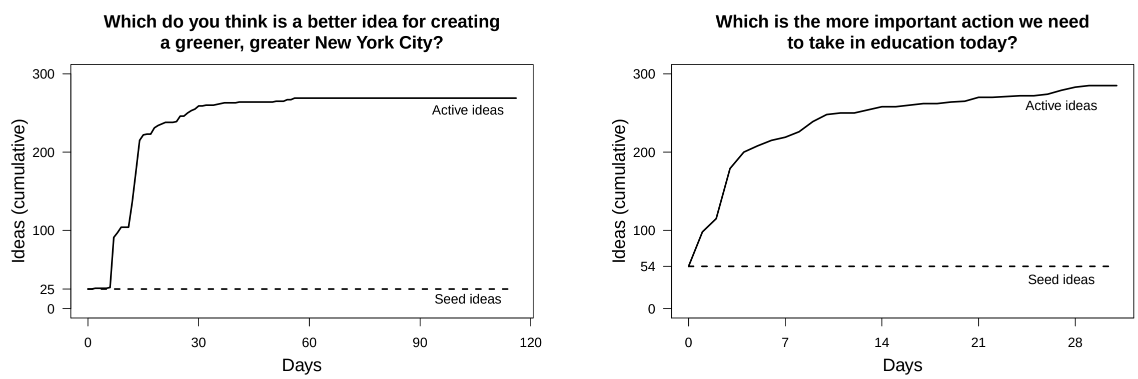
Patterns of respondent participation in the two wiki surveys were also quite similar. Within each survey, levels of respondent contribution varied widely, in terms of both number of votes cast and number of new ideas contributed. In both cases, the distributions of both votes and idea uploads contained "fat heads" and "long tails" (see Figs. 6(a) and 6(b)). If the wiki surveys captured



(a) PlaNYC

(b) OECD

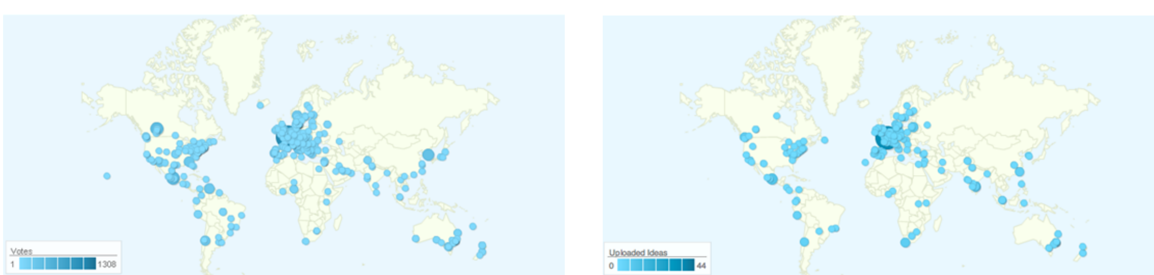
Figure 3: Cumulative number of votes in each wiki survey.



(a) PlaNYC

(b) OECD

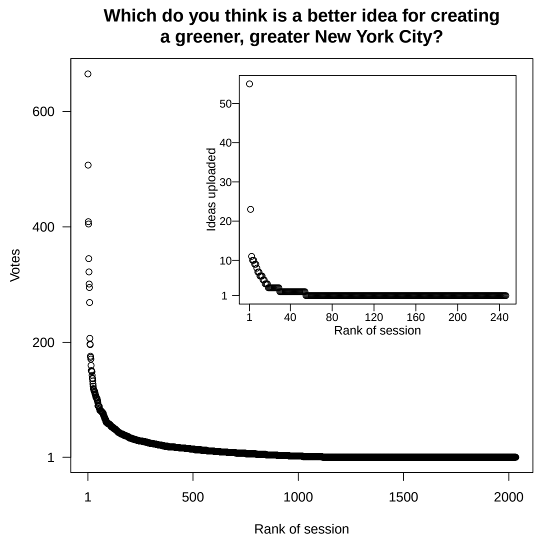
Figure 4: Cumulative number of activated ideas. In both cases the pool of ideas grew over time as participants contributed to the wiki survey. PlaNYC had 25 seed ideas and 464 user-submitted ideas, 244 of which the Mayor's Office activated. The OECD had 54 seed and 533 user-submitted ideas, 231 of which it activated. In both cases, ideas that were deemed inappropriate or duplicative were not activated.



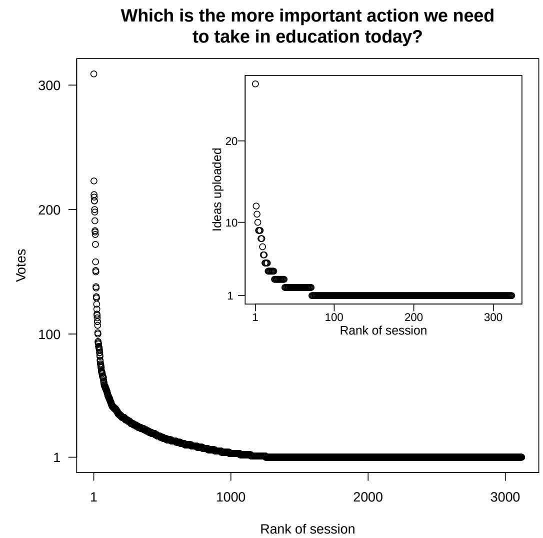
(a) Votes

(b) Uploaded Ideas

Figure 5: Geographic distribution of voting and uploading of ideas in the OECD wiki survey. Votes were cast from more than 90 countries and ideas submitted from more than 50 countries.



(a) PlaNYC



(b) OECD

Figure 6: Both the number of votes per session and the number of uploads per session show a "fat head" and a "long tail." Note that the scales on the figures are different.

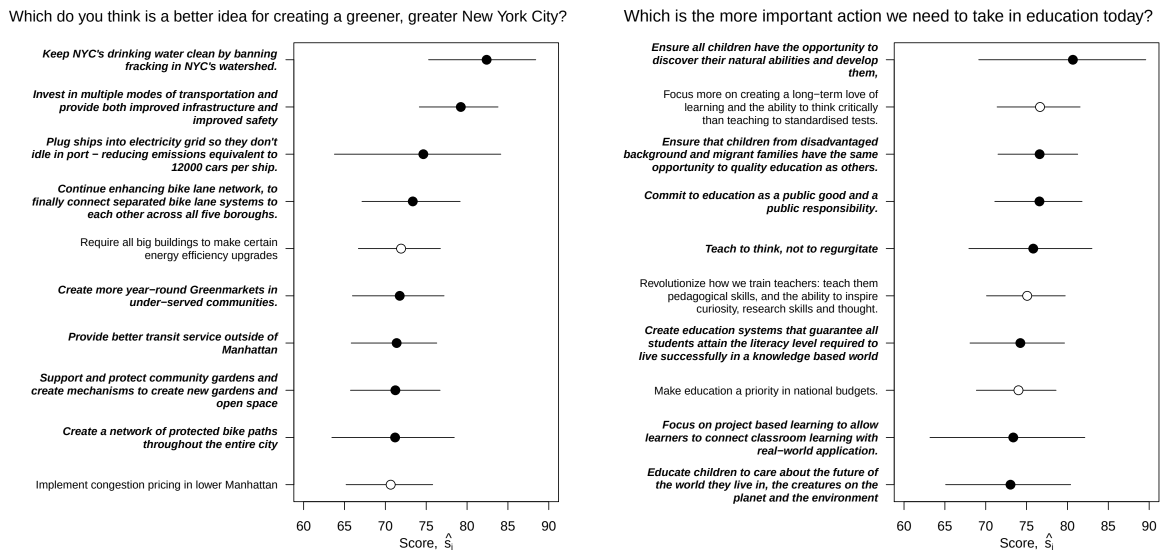
only a fixed amount of information per user session—as opposed to capturing all levels of effort—a significant amount of information would have been lost; for instance, if 10 and only 10 votes from each session were captured, approximately 75% of the votes collected in each survey would have been discarded. Further, if we were to limit the number of ideas uploaded to one per session, as is typical in surveys with one and only one "other box," we would have excluded a significant number of user-submitted new ideas: nearly half of the user-submitted ideas in the PlaNYC survey[6] and about 40% in the OECD survey.

Further, in both wiki surveys, many of the highest scoring ideas were uploaded by users. For PlaNYC, 8 of the top 10 ideas were uploaded by users, as were 7 of the top 10 for the OECD (Fig. 7). This finding is clear evidence of the importance of the collaborative aspect of wiki surveys; these uploaded ideas would never have been captured by a traditional survey. We note, however, that not all uploaded ideas were high-scoring: the scores of uploaded ideas had higher variance than the scores of seed ideas (Fig. 8). In other words, while some uploaded ideas were very popular, many were quite unpopular. Since many wiki survey creators are primarily looking to find the best ideas, the high variability of the uploaded ideas is an asset, not a liability.

6.2 Contextualizing the results

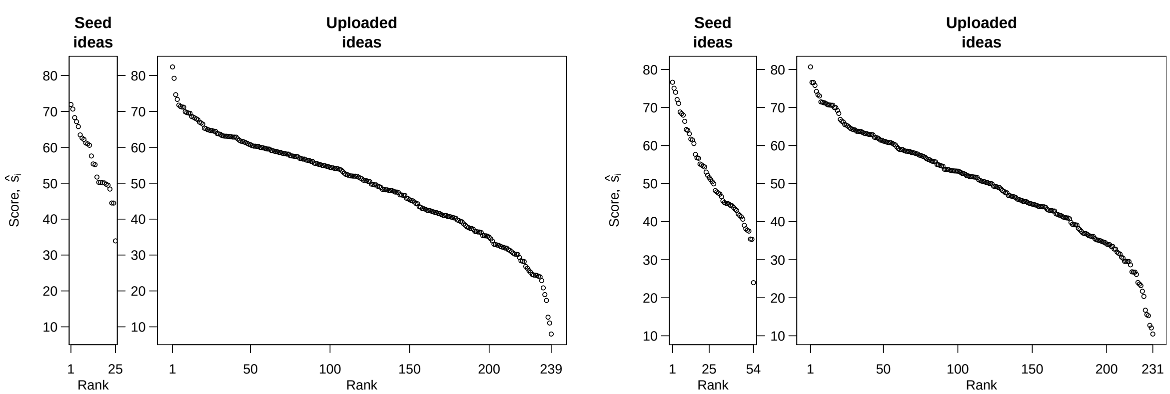
In order to better understand what kinds of uploaded ideas scored well, we conducted semi-structured interviews with officials at the OECD and the New York City Mayor's Office. Based on these interviews, as well as interviews with six wiki survey creators from other groups, there seem to be two general classes of uploaded ideas that score well: *novel information*—that is, new ideas

[6] In this case, the number of uploaded ideas per session is somewhat difficult to interpret because some user-submitted ideas were bulk-uploaded by the Mayor's Office following community meetings at which ideas were recorded on paper. Unfortunately no records were kept of this bulk uploading, so we cannot distinguish it from other user behavior.



(a) PlaNYC (b) OECD

Figure 7: Top 10 ideas (with at least 50 completed appearances). Ideas that were uploaded by users 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 uploaded by users. In the case of the OECD, 7 of the 10 highest scoring ideas were uploaded by users. Horizontal lines show 95% posterior intervals.



(a) PlaNYC (b) OECD

Figure 8: Distribution of scores of seed ideas and user-submitted ideas (with at least 50 completed appearances). In both cases, some of the lowest-scoring ideas were user-submitted, but critically, some of the highest-scoring ideas were also submitted by users. In general, the large number of user-submitted ideas, combined with their high variance, means that they typical include some extremely popular ideas. Posterior intervals for each estimate are not shown.

that were not anticipated by the wiki survey creators—and *alternative framings*—that is, new and "stickier" ways of expressing existing ideas.

An example of one of these alternative framings was one of the top-scoring (and user-uploaded) ideas in the OECD wiki survey: "Teach to think, not to regurgitate." While this idea may not be substantively new, our interviews with the OECD suggest that the wiki survey respondent who contributed it expressed it in a way that the organization might not have:

> "['teach to think, not to regurgitate'] . . . it wouldn't be formulated in such a way [by the OECD]. So that's probably the most, it's not controversial, but it's the most surprising, it's very un-OECD-speak, which we liked." (interview with Julie Harris, OECD, February 3, 2011)

In the OECD context, the top-scoring ideas tended to be phrased as aspirational principles (e.g., "ensure that children from disadvantaged background[s] and migrant families have the same opportunity to quality education as others"). Our interviews revealed that the success of ideas framed in this way was informative for the organization's staff:

> "In terms of substance, what for me has been most interesting is that . . . those top priorities . . . they're all very much couched in the language of principles, basic principles. . . . It's sort of constitutional language." (interview with Joanne Caddy, OECD, February 15, 2011)

PlaNYC's wiki survey creators noted similar alternative framings. The top-scoring idea ("Keep NYC's drinking water clean by banning fracking in NYC's watershed")[7] was submitted by a user; policy advisor Ibrahim Abdul-Matin reported that although the Mayor's Office strongly supports a ban on fracking, it would not have phrased the initiative as such, but, instead, would have used more general language. Seeing the amount of support for a fracking ban specifically framed in this way enables the Mayor's Office to understand how to frame PlaNYC's initiatives so that they are more accessible to and resonant with New York City's residents:

> "We hate fracking, and we've been very clear about it. And so this for us, this says, we have support amongst the people who are paying attention to be very clear that we don't like fracking. [Question: Okay. So why was fracking not one of the seed ideas that you guys had?] Because we talk about it differently. We'll say, 'protect the watershed.' We don't say, 'protect the watershed from fracking.' " (interview with Ibrahim Abdul-Matin, Policy Advisor, December 12, 2010)

More generally, Abdul-Matin noted the value of information about what framings were most resonant with the public:

> "I think the value of this process has been to find new ways to talk about and explain what we're doing. It's a framing challenge. ... Because really the biggest challenge with PlaNYC 2.0 is making sure that it has value and resonance for people who actually live in the city, so it's not just some abstract plan." (interview with Ibrahim Abdul-Matin, Policy Advisor, December 12, 2010)

In addition to alternative framings, other high-scoring user-uploaded ideas contained information that was new to the wiki survey creator. For example, in the PlaNYC context, the Mayor's

[7] "Fracking," short for hydraulic fracturing, is a drilling technique for extracting oil and natural gas.

Office reported that user-submitted ideas were sometimes able to bridge multiple policy arenas (or "silos") that might have been more difficult connections to make for people working within a specific arena. Consider the high-scoring user-submitted idea "plug ships into electricity grid so they don't idle in port—reducing emissions equivalent to 12000 cars per ship," an idea that bridges multiple policy arenas. A policy advisor suggested that as such, staff may not have prioritized such an idea internally (it did not appear on the Mayor's Offices list of "seed" ideas), but that public support for it suggests its importance as a 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 emissions, reducing emissions, is something that's an overall goal of the plan. We want to reduce carbon emissions 30 percent by 2030. So this has a lot of value to it for us to learn from." (interview with Ibrahim Abdul-Matin, Policy Advisor, December 12, 2010)

Taken together, these case studies suggest that wiki surveys may provide information that is difficult, if not impossible, to gather from a more traditional survey instrument. This unique information may involve both the content of the ideas that are submitted by users and the language used to frame them.

7 Limitations and future research

These wiki surveys, as well as the more than 1,500 others that have been created, show that pairwise wiki surveys are a viable method for social data collection, but clearly there are also limitations that need to be carefully considered. We are optimistic, however, that many of these challenges can be overcome. Just as the shift from face-to-face surveys to telephone surveys created both research opportunities and methodological challenges (Groves, 1990; Brick and Tucker, 2007), so too would a shift from traditional surveys to wiki surveys. In this spirit, we see wiki surveys not as a break from traditional survey methods, but as part of the long evolution of the field in response to new opportunities created by changes in technology and society (Mitofsky, 1989; Dillman, 2002; Couper, 2011; Newport, 2011; Groves, 2011).

In this paper we have focused on the score of each item (Eq. 6), and a natural next step would be to estimate additional information about the contours of public opinion from the votes. New York City, for example, might wish to know how residents of the city's five boroughs differ in how they think the city should become "greener and greater." We do not currently collect demographic information about the voter, so we cannot perform such an analysis.[8] However, if this information were collected, these co-variates at the level of the user-session could be added into the modeling framework described in Section 5. Additional information about public opinion could also be estimated by collecting co-variates about each idea—for example, whether the idea has to do with bicycling, public transportation, etc. This would allow researchers to estimate the broad outlines of public opinion, rather than support for specific items.

Further, even though the results of the wiki surveys were informative to the creators, as described in our interviews, one may wonder whether the wiki surveys accurately reflect public opinion. Ultimately, this is an empirical question, albeit one that is not yet resolved. Validation studies

[8]We opted not to collect demographic information in order to minimize barriers to participation that might create differential non-response (Groves et al., 1992).

should be run to check, among other things, how robust our results are to the *ad hoc* sampling approaches used by the groups in the case studies and to check the predictions from the model.

Another question for further research is how we make the estimates described above more efficiently—either by requiring fewer votes or fewer participants. Our Bayesian framework provides a natural approach to this problem: we should choose the prompt that tightens the posterior distribution most for the parameters of interest (Lindley, 1956; Glickman and Jensen, 2005), a solution suggested by the literature on Bayesian experimental design (Chaloner and Verdinelli, 1995). However, maximizing the amount of information per vote may not maximize the amount of information per participant, which is determined by both the information per vote and the number of votes cast (von Ahn and Dabbish, 2008). That is, an algorithm that chooses very informative prompts from a statistical perspective might not be effective if people do not enjoy responding to those kinds of prompts. Thus, algorithms need to address both maximization of information per prompt and participation by, for example, choosing prompts (or sequences of prompts) that respondents enjoy answering.

Finally, we note that pairwise wiki surveys are best suited for situations in which there is a single predetermined question. Future research might develop methods for mixing elements of pairwise wiki surveys with traditional surveys. Additional types of wiki surveys could also be developed that are amenable to multiple predetermined questions, or even in which respondents themselves help to generate the questions (Sullivan et al., 1979; Gal and Rucker, 2011).

Given the numerous limitations described above, we have taken several steps to facilitate the rapid development of wiki surveys. First, we made it easy for users of `allourideas.org` to download the data from their wiki survey for offline analysis. We hope this will lead to new ways of estimating information from the votes and will make the website a useful tool for others. Second, we have created an Application Programming Interface (API) which allows anyone to build a pairwise comparison website using the tools and algorithms that we have developed. Finally, all the code that powers the API and `allourideas.org` is available open-source to anyone who wishes to use or improve it (`github.com/allourideas`). We hope that these steps will lead to rapid development of wiki surveys so that researchers can take full advantage of the possibilities of this open and quantifiable form of data collection.

A Website implementation

When implementing pairwise wiki surveys at `www.allourideas.org`, we encountered two main methodological issues: 1) choosing prompts to present to voters and 2) using the votes to estimate public opinion. We solved these problems using relatively simple heuristic approaches that were able to run in real time. Because these approaches were used during our data collection, we present them here for completeness. However, in future research, as described in Section 7, we plan to unify these two issues by displaying subsequent prompts that maximize the information learned about the most important parameters in the model (Chaloner and Verdinelli, 1995).

A.1 Selection of prompts

The simplest way to select prompts for the voters would be to sample with uniform probability from the set of prompts. However, because the wiki surveys are collaborative, participants contribute new items throughout the voting process so prompts with user-submitted items will tend to have fewer completed contests (a completed contest occurs when a prompt is sent to a voter and the voter returns a choice of one item or the other). This means that using the simplest approach would result in more completed contests—and therefore more precise estimates—for seed items than user-submitted items. This is far from ideal because the user-submitted items are potentially the most interesting ones. Instead, it is preferable to spread the appearances more evenly over the set of prompts. Therefore, we developed a "catch up" algorithm, which shows prompts with fewer completed appearances with higher probability. In essence, it helps newer prompts "catch up" to older ones in terms of number of appearances. Specifically, the draw-wise probability for a given prompt (i, j) is:

$$p_{i,j} = \frac{min\left(\frac{\frac{1}{(n_{i,j}+1)^{\alpha}}}{c_1}, \tau\right)}{c_2} \tag{7}$$

where $n_{i,j}$ is the number of completed contests for prompt (i, j), α is a parameter that tunes the weight of the number of completed contests, and τ is a "throttle" to ensure that the draw-wise probability never exceeds some threshold (it could create a poor user experience if the same prompt had a draw-wise probability of, say, 0.5). Finally, c_1 and c_2 are normalizing constants to ensure that the distribution sums to 1.[9] Although somewhat awkward-looking, Eq. 7 is straightforward to implement and runs very quickly. As a first step we choose $\alpha = 1$ and $\tau = 0.05$, but the optimal values of these parameters are an open question.

A.2 Estimating public opinion

The estimate of public opinion that is presented to users is what we call the score of each item (see Fig. 2). The score of an item is the probability that the item will beat a randomly chosen item for a randomly chosen user-session. Given this focus on the probability of a win, a binomial model seems appropriate. If one assumes a uniform prior for a binomial random variable, the resulting posterior for the probability of a win follows a Beta distribution (Hoff, 2009). If we multiply the expected value of that Beta distribution by 100 (to place things on a more natural scale), we have

$$\widehat{s_i'} = \frac{(w_i + 1)}{(w_i + 1) + (l_i + 1)} \times 100 \tag{8}$$

[9]The normalizing constants are $c_1 = \sum_i \frac{1}{(n_{i,j}+1)^{\alpha}}$ and $c_2 = \sum_i min\left(\frac{\frac{1}{(n_{i,j}+1)^{\alpha}}}{c_1}, \tau\right)$ where τ is the throttle, the maximum probability for a pair appearing in a draw.

where w_i is the number of wins for item i and l_i is the number of losses for item i; see Hoff (2009, Ch. 3) for a derivation. Thus, the estimated score ranges from 0 to 100 and resembles a simple winning percentage with an additional term that provides some smoothing.

This approach is both easy to calculate and reasonably principled because it is derived from standard Bayesian methods. It also has several nice practical properties including that it produces a reasonable estimate for new items that have not appeared (e.g., $\widehat{s}_i' = 50$) and the amount the score changes with any specific vote decreases as the number of votes on the item increase. However, the approach also has some limitations. First, it does not account for the fact that votes are nested within voters. In other words, a person who votes 100 times will have 100 times the influence as someone who votes only once. Also, this approach does not consider the "strength of schedule" (i.e., the items that a given item has competed against). For example, this scoring approach gives equal weight to an item beating a popular item as to one beating an unpopular item. For these reasons and others, we developed the model described in Section 5 of the paper, which does not suffer from these two limitations, but which takes many hours to compute. In the cases considered in the paper, however, both the simple score (Eq. 8) and the modeled score (Eq. 6) turned out to be quite similar (correlation of greater than 0.95 in both cases). In the future, we hope to calculate the modeled score (Eq. 6) on the actual website; we suspect that approaches using variational inference will make that possible (see, e.g., Braun and McAuliffe (2010)).

B Model and computation

As described in the main text, a main statistical challenge is to use the votes (an example set of votes is listed in Table 1) to estimate public opinion, which is this case is defined to be a matrix $\mathbf{\Theta}$ that represents how much the user in each session supports each idea. We begin by assuming a model for how the votes are generated, and a natural first choice would be

$$Pr(a \text{ beats } b \text{ in session } j) = F(\theta_{j,a} - \theta_{j,b}) \tag{9}$$

where $\theta_{j,a}$ is the appeal of item a to the person in session j. That is, the probability that item a beats item b is a function of the difference in the appeals of the two items $\theta_{j,a}$ and $\theta_{j,b}$. In previous work, numerous functional forms have been assumed for F, but the two common choices are the cumulative standard normal—resulting in the Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 1951)—or the logistic function—leading to the Bradley-Terry model (Bradley and Terry, 1952). In fact, Stern (1990) has shown that the Thurstone-Mosteller model and the Bradley-Terry model can both be viewed as special cases of a more general model, and empirically, both models produce estimates that are essentially equivalent (Stern, 1992). However, the Thurstone-Mosteller model is much more efficient computationally because it more naturally leads to the Gibbs sampling updates as described below. For that reason we assume that

$$Pr(a \text{ beats } b \text{ in session } j) = \Phi(\theta_{j,a} - \theta_{j,b}) \tag{10}$$

where Φ is the cumulative standard normal distribution. Thus, we map the difference between the appeals, which ranges from $-\infty$ to ∞, to a value that ranges from 0 to 1. Future work could explore the robustness of our estimates to the choice of the standard normal or could attempt to estimate the shape of F directly. Another extension of the model would allow for ties, which are not allowed in our modeling framework. For more on ties in pairwise comparison models see David (1988).

Given the voting model described in equation 10, the likelihood can be written to resemble a

Vote ID	User-session ID	Prompt	
1	1	**item 1**	item 4
2	1	item 3	**item 1**
3	1	**item 4**	item 3
4	2	**item 3**	item 4
5	2	item 4	**item 2**

Table 1: An example of five responses given in two user-sessions. The bolded item is the one that was chosen by the respondent.

standard probit model, given a properly constructed design matrix $\boldsymbol{X}$ and outcome vector $\boldsymbol{Y}$,

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \tag{11}$$

In this case, $\boldsymbol{X}$ has V rows and $J \times K$ columns, where V is the number of votes, J is the number of user-sessions, and K is the number of items. Therefore, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{im})$ and $m = J \times K$. In order for the algebra to work out properly, each row in $\boldsymbol{X}$ has a "1" in the column of the user-session/item that appeared on the left of the pair and a "-1" on the column of the user-session/item that appeared on the right of the pair. $\boldsymbol{Y}$ is a vector with V entries that has a "1" if the item on the left wins and "0" if the item on the right wins. For example, the votes in Table 1 would lead to

$$\boldsymbol{Y} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{array}{c} \begin{array}{cccccccc} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} & \theta_{1,4} & \theta_{2,1} & \theta_{2,2} & \theta_{2,3} & \theta_{2,4} \end{array} \\ \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & -1 & & & & \\ -1 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & -1 & 1 & & & & \\ \hline & & & & 0 & 0 & 1 & -1 \\ & 0 & & & 0 & -1 & 0 & 1 \end{array} \right) \end{array}$$

By explicitly estimating how much each person likes each item, this modeling approach naturally allows for heterogeneity in the preferences of the respondents. However, the cost of such flexibility is that there are an enormous number of parameters to be estimated; in the case of PlaNYC, there were about 500,000 parameters to estimate (∼2,000 user-sessions × ∼250 items) and in the case of the OECD there are more than 800,000 parameters (∼3,000 user-sessions × ∼275 items). Therefore, to add more structure to the problem and to allow for partial pooling of information across respondents (Gelman and Hill, 2006; Rossi et al., 2006), we add hierarchical terms in the model that assume that the preferences for each item are normally distributed with an item specific mean μ_k and a common standard deviation of σ,

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\mu}, \sigma) = \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \times \prod_{j=1}^{J} \prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \tag{12}$$

where $\boldsymbol{\mu} = \mu_1 \ldots \mu_K$ and σ is assumed to be 1.[10] Future work could improve the model by estimating σ directly, estimating a σ_k for each item, or even estimating the functional form that the $\theta_{j,k}$ follow for each k.

[10] As a robustness check, we re-ran the model with $\sigma = 0.5$ and $\sigma = 2$. In both cases, the results were essentially the same as when $\sigma = 1$.

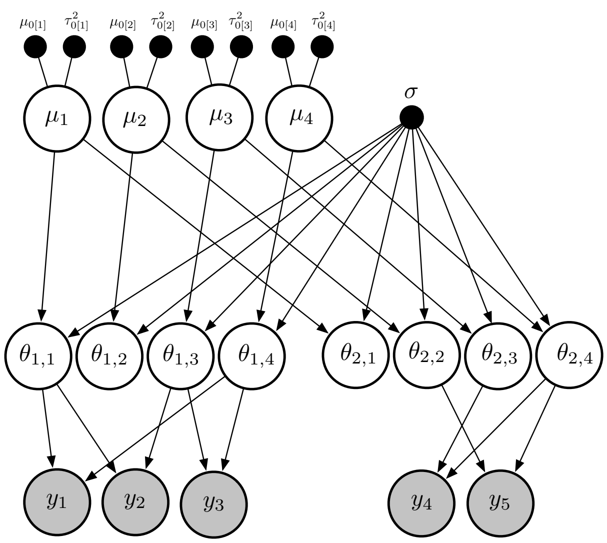


Figure 9: Graphical representation of the model (see Equation 13). This graphical model shows the assumed data generating process for the sample data shown in Table 1. At the top of the figure, priors are used to generate item-specific means (e.g., μ_1). Next, these parameters and σ, which is assumed to be 1, are used to generate the elements of the public opinion matrix (e.g., $\theta_{1,1}$). Finally, these elements of the public opinion matrix generate the observed votes (e.g., y_1). The challenge is then to estimate the unknown parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\mu}$) from the observed data.

Finally we add conjugate priors to yield the following posterior distribution:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\mu}, \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}) \quad \propto \quad & \prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta}))^{1-y_i} \times \prod_{j=1}^{J} \prod_{k=1}^{K} N(\theta_{j,k} \mid \mu_k, \sigma) \\ & \times \prod_{k=1}^{K} N(\mu_k \mid \mu_{0[k]}, \tau^2_{0[k]}) \end{aligned} \tag{13}$$

The model is represented graphically in Figure 9.

As is common in discrete-choice models (Train, 2009), the model above is only weakly identified because one could add a constant c to all the θ parameters and leave the posterior largely unchanged.[11] Therefore, we pick an arbitrary item to have $\mu_k = 0$ which requires setting the hyper-parameters $\mu_{0[k]} = 0$ and $\tau^2_{0[k]} = 0.000001$. For the remaining items, we set weakly informative priors: $\mu_{0[k]} = 0$, $\tau^2_{0[k]} = 4$.

Before attempting to sample from this posterior distribution, we perform two "tricks" that greatly facilitate computation, but which do not affect the underlying model that we are estimating. First, we note that in practice many user-sessions do not encounter many of the items. For example, in the voting data in Table 1, user-session 1 never encountered item 2 and user-session 2 never encountered item 1. Thus, we note that there are actually two types of θ parameters, those that are informed by a specific vote (in this case, $\theta_{1,1}$, $\theta_{1,3}$, $\theta_{1,4}$, $\theta_{2,2}$, $\theta_{2,3}$, $\theta_{2,4}$) and those that are not (in this case, $\theta_{1,2}$, $\theta_{2,1}$). Thus, we note that

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\mu}, \sigma) = p(\boldsymbol{\theta_v} \mid \boldsymbol{Y}, \dot{\boldsymbol{X}}, \boldsymbol{\mu}, \sigma) \times p(\boldsymbol{\theta_h} \mid \boldsymbol{\mu}, \sigma) \tag{14}$$

[11] It is perhaps easier to see this non-identifiability in the model for a single vote (Eq. 10).

where $\boldsymbol{\theta_v}$ are parameters that are estimated from the votes and the hyper-parameters and $\boldsymbol{\theta_h}$ are parameters that depend on the votes only through the hyper-parameters, and $\dot{\boldsymbol{X}}$ is the reduced form of the original design matrix $\boldsymbol{X}$ that only includes columns for $\theta \in \boldsymbol{\theta_v}$. For example. for the votes in Table 1, $\dot{\boldsymbol{X}}$ is

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \theta_{1,1} & \theta_{1,3} & \theta_{1,4} & \theta_{2,2} & \theta_{2,3} & \theta_{2,4} \end{array} \\
\dot{\boldsymbol{X}} = & \left(\begin{array}{ccc|ccc}
1 & 0 & -1 & & & \\
-1 & 1 & 0 & & 0 & \\
0 & -1 & 1 & & & \\
\hline
 & & & 0 & 1 & -1 \\
 & 0 & & -1 & 0 & 1
\end{array}\right)
\end{array}
$$

In this simple example $\dot{\boldsymbol{X}}$ is 33% smaller than $\boldsymbol{X}$, but in the cases considered in the paper the reduction is much more substantial: In the PlaNYC pairwise wiki survey, $\dot{\boldsymbol{X}}$ is about 90% smaller than $\boldsymbol{X}$, and in the OECD pairwise wiki survey, $\dot{\boldsymbol{X}}$ is about 95% smaller than $\boldsymbol{X}$. Reducing the size of the design matrix in this way yields a substantial savings in terms of computing time and RAM needed to approximate the posterior. Given this trick, we can re-write equation 12 as follows:

$$
\begin{aligned}
p(\boldsymbol{\theta_v}, \boldsymbol{\theta_h} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\mu}) \propto \quad & \left(\prod_{i=1}^{V} \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta_v})^{y_i} (1 - \Phi(\boldsymbol{x}_i^T \boldsymbol{\theta_v}))^{1-y_i} \times \prod_{(j,k)}^{\theta_{j,k} \in \boldsymbol{\theta_v}} N(\theta_{j,k} \mid \mu_k, \sigma) \right) \\
& \times \left(1 \times \prod_{(j,k)}^{\theta_{j,k} \in \boldsymbol{\theta_h}} N(\theta_{j,k} \mid \mu_k, \sigma) \right)
\end{aligned} \tag{15}
$$

A second computational trick is to note that by introducing a latent variable $\boldsymbol{z}$ we are able to sample from the posterior more easily, an approach sometimes called data augmentation. Roughly, we are assuming that although we observe a discrete outcome y_i, there is actually an underlying continuous value z_i that generates y_i. As shown by Albert and Chib (1993), including this continuous latent variable in our model enables us to sample from the posterior distribution more easily. For a more thorough discussion of this type of data augmentation, see Lynch (2007) and Jackman (2009).

Combining these two "tricks," we are left with the following posterior distribution:

$$
\begin{aligned}
& p(\boldsymbol{\theta_v}, \boldsymbol{\theta_h}, \boldsymbol{z}, \boldsymbol{\mu} \mid \boldsymbol{Y}, \dot{\boldsymbol{X}}, \sigma, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}) \propto \\
& \quad \left(\prod_{v=1}^{V} (I(z_i > 0) I(y_i = 1) + I(z_i < 0) I(y_i = 0)) \times N(z_i \mid \dot{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{\theta_v}, 1) \times \prod_{(j,k)}^{\theta_{j,k} \in \boldsymbol{\theta_v}} N(\theta_{j,k} \mid \mu_k, \sigma) \right) \\
& \quad \times \left(1 \times \prod_{(j,k)}^{\theta_{j,k} \in \boldsymbol{\theta_h}} N(\theta_{j,k} \mid \mu_k, \sigma) \right) \times \prod_{k=1}^{K} N\left(\mu_k \mid \mu_{0[k]}, \tau_{0[k]}^2 \right)
\end{aligned} \tag{16}
$$

To make draws from this posterior distribution in equation 16 we use Markov chain Monte Carlo, specifically Gibbs sampling (Gelfand and Smith, 1990). That is, we repeatedly draw from the conditional distribution for each parameter given the current values of the other parameters; for a review of Gibbs sampling, see Gelman et al. (2003).[12] As is standard practice (Gelman and Shirley, 2011), we ran three parallel chains from over-dispersed starting values for 200,000 steps, saving every

[12] Alternative approaches using variational inference (Braun and McAuliffe, 2010) represent an important possible extension of this approach, especially as the number of parameters to be estimated increases.

200th draw, and discarded the first half of each chain as burn-in. At that point, all parameter estimates had approximately converged, $\hat{R} < 1.1$ (Gelman et al., 2003), and so we combined the post burn-in draws to summarize the posterior distribution. Overall, these computations took about 36 hours per dataset on a fast desktop computer. Computer code to make these draws was written in `R` (R Development Core Team, 2011) and utilized several packages: `plyr` (Wickham, 2011a), `multicore` (Urbanek, 2011), `bigmemory` (Kane and Emerson, 2011), `truncnorm` (Trautmann et al., 2011), `testthat` (Wickham, 2011b), and `Matrix` (Bates and Maechler, 2011). Upon publication of this article, all code used to produce analysis in this paper will be released to the public.

When fitting the model to the case studies in this paper, we estimated parameters for items that were active at the end of voting and had at least one win and at least one loss. Votes involving ideas that did not meet these criteria were not used in model fitting. We also did not use votes that were flagged by the website as being invalid.[13] To find this set of items to estimate and votes to use, we began with all raw votes (PlaNYC: 28,829, OECD: 28,599) and all ideas that were active at the end of the wiki survey (PlaNYC: 269, OECD: 285). After excluding invalid votes, we were left with a proposed set of votes to use for estimation (PlaNYC: 26,728, OECD: 25,475).[14] Next, we created a proposed set of ideas to estimate from all ideas that were active at the end of voting and had at least one win and one loss in the proposed set of votes (PlaNYC: 269, OECD: 285). Then we dropped all votes that were not between two ideas in the proposed set of ideas resulting a new proposed set of votes for estimation (PlaNYC: 26,644, OECD: 23,922). Finally, we verified that given the proposed set of votes for estimation and the proposed set of ideas for estimation, all ideas to be estimated had at least one win and at least one loss. In the end, for PlaNYC this procedure yielded 26,644 votes on 269 ideas cast in 2,016 user-sessions, and for the OECD this procedure yielded 23,922 votes on 285 ideas cast in 2,851 user-sessions.

These votes and ideas were then used to fit the model in equation 16 using Gibbs sampling with five update steps.

• Step 1: Draw $\boldsymbol{z} \mid \boldsymbol{Y}, \boldsymbol{\theta_v}, \dot{\boldsymbol{X}}$

Recall that $\boldsymbol{z}$ is the underlying latent outcome which we cannot observe. Based on ideas developed in Albert and Chib (1993), we sample $\boldsymbol{z}$ from a truncated normal distribution such that $z_i > 0$ if $y_i = 1$ and $z_i < 0$ if $y_i = 0$. That is,

$$z_i \sim \begin{cases} N(\dot{\boldsymbol{x}}_i^T \boldsymbol{\theta_v}, 1) I(z_i^* > 0) & \text{if } y_i = 1 \\ N(\dot{\boldsymbol{x}}_i^T \boldsymbol{\theta_v}, 1) I(z_i^* < 0) & \text{if } y_i = 0 \end{cases} \tag{17}$$

where I is an indication function which equals 1 when its argument is true and 0 when false (Jackman, 2009). This indicator function ensures that we are drawing from a properly truncated dis-

[13]Flagging some votes as invalid is part of our attempt to make the wiki survey more manipulation-resistant (Resnick and Sami, 2008). In all data collection, researchers must be wary of respondents who wish to manipulate results, but those risks are probably higher than typical in this research. There are two ways that a vote can be flagged as invalid. First, if we receive multiple responses for the same prompt appearance (as would occur if the respondent tried to click several times before the page reloads), then only the first response is marked valid; the others are marked invalid. Second, all votes that occur immediately following a response "I can't decide" (see voting screen in Fig. 2) are marked invalid. These votes are not included in estimation because in a previous wiki survey we detected a respondent who attempted to manipulate the results by clicking "I can't decide" until his or her preferred idea was presented, at which point he or she voted for that idea. Our flagging procedure prevents this manipulation from influencing the results. Though our approach probably invalidates some legitimate data, we prefer to err on the side of caution. Finally, we note that these procedures do not protect against all possible forms of manipulation, and future research will be necessary to make wiki surveys more manipulation-resistant. In the two case studies presented in this paper, we do not believe that any large manipulations took place.

[14]The OECD wiki survey had a period of internal pilot testing from September 3, 2010 to September 15, 2010. All votes cast during this time were flagged as invalid and all ideas uploaded during this time were set to be seed ideas.

tribution. Computationally, we draw from the truncated normal using the `truncnorm` package in `R` (Trautmann et al., 2011).

• Step 2: Draw $\boldsymbol{\theta_v} \mid \boldsymbol{z}, \boldsymbol{\mu}, \dot{\boldsymbol{X}}, \boldsymbol{\sigma}$

Under the data augmentation approach of Albert and Chib (1993), once we have simulated $\boldsymbol{z}$, the latent outcome, we are left with a standard hierarchical linear model. To update $\boldsymbol{\theta_v}$ we use the "all-at-once" approach described in Gelman et al. (2008).

That is,

$$\boldsymbol{\theta_v} \sim N(\hat{\boldsymbol{\theta}}_{\boldsymbol{d}}, \hat{\boldsymbol{V}}_{\boldsymbol{\theta_v}}) \tag{18}$$

where

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{d}} = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{Y}} \quad , \quad \hat{\boldsymbol{V}}_{\boldsymbol{\theta_v}} = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{X}})^{-1}$$

$$\widetilde{\boldsymbol{X}} = \begin{pmatrix} \dot{\boldsymbol{X}} \\ \boldsymbol{I} \end{pmatrix} \quad , \quad \widetilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ \tilde{\boldsymbol{\mu}} \end{pmatrix} \quad \text{, and} \quad \widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma_y} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_\theta} \end{pmatrix}.$$

Further, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{\Sigma_y} = \text{Diag}(1)$, $\boldsymbol{\Sigma_\theta} = \text{Diag}(\sigma)$, and $\widetilde{\boldsymbol{\mu}}$ is a vector that is the same length as $\boldsymbol{\theta_v}$ and represents an "expanded" version of $\boldsymbol{\mu}$. That is, if the i^{th} column of $\dot{\boldsymbol{X}}$ represents item k (independent of what user-session is involved), then the i^{th} element of $\widetilde{\boldsymbol{\mu}}$ is μ_k.

Computationally, we note that $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are almost all zeros, so the calculations described above to make a draw are made using sparse matrix routines that are implemented in the `Matrix` package in `R` (Bates and Maechler, 2011).

• Step 3: Update $\boldsymbol{\theta_h} \mid \boldsymbol{\mu}, \sigma$

A large number of the θ parameters are determined by data only through the hyper-parameters. For these θ, which we call $\boldsymbol{\theta_h}$, we update as follows:

$$\theta_{j,k} \sim N(\mu_k, \sigma) \quad \forall \quad \theta_{j,k} \in \boldsymbol{\theta_h} \tag{19}$$

Thus, this step is roughly like an imputation based on the overall estimated characteristics of the population. Computationally, no special steps are required to make these updates.

We note that this step highlights an assumption that we are making about the preferences of our respondents: that the θ's in $\boldsymbol{\theta_v}$ are directly informative about the θ's in $\boldsymbol{\theta_h}$. We can think of two cases in which this might not be true. First, consider an item uploaded by a person in user-session j. All user-sessions before j did not have a chance to vote on this item so we will estimate their liking of this item based on the users after user-session j. Therefore, if for some reason the preferences of users to the site vary systematically over time, this procedure will not work well. Second, the greedy nature of the wiki survey could also lead to problems if people who vote many times have systematically different preferences than those who vote fewer times. For example, imagine that there are two types of people: vegans and non-vegans. Further, imagine that all vegans love bicycles, all non-vegans hate bicycles, and that vegans cast more votes than non-vegans. Now, if we have a user-session j that did not vote on an idea k ("more bike racks in Manhattan"), the model will estimate $\theta_{j,k}$ based on the other $\theta_{\cdot,k} \in \boldsymbol{\theta_v}$. But, in this case, the $\theta_{\cdot,k} \in \boldsymbol{\theta_v}$ over-represent opinions of vegans relative to non-vegans. This example shows that an important extension to the model would include co-variates at the level of the user-session and at the level of the item, not just because these are substantively meaningful, but because these can reduce distortions caused by the unequal amount of votes that we have from each user-session. We think these issues of model robustness will be an important area of future research.

• **Step 4:** Update $\boldsymbol{\mu} \mid \boldsymbol{\theta_v}, \boldsymbol{\theta_h}, \sigma, \boldsymbol{\mu_0}, \boldsymbol{\tau_0^2}$

$$\mu_k \sim N(\mu, \tau^2) \tag{20}$$

where

$$\mu = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{\theta}_{\cdot,k}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \tau^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

where $\bar{\theta}_{\cdot,k}$ is the mean of the θ for a specific item k (that is, $\frac{1}{J}\sum_{j=1}^{J}\theta_{j,k}$) and n is the number of estimates involved (in this case, the number of user-sessions, J). See Hoff (2009, Ch. 6) for a derivation. No special computational issues are involved in this update.

References

Albert, James H. and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. New York, NY: Hyperion.

Balasubramanian, Siva K. and Wagner A. Kamakura. 1989. "Measuring Consumer Attitudes Toward the Marketplace with Tailored Interviews." *Journal of Marketing Research* 26:311–326.

Bates, Douglas and Martin Maechler. 2011. "Matrix: Sparse and Dense Matrix Classes and Methods." R package version 0.999375-50.

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.

Bockenholt, Ulf. 2007. "Thurstonian-based Analyses: Past, Present, and Future Utilities." *Psychometrika* 71:615–629.

Bradley, Ralph A. and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39:324–345.

Braun, Michael and Jon D. McAuliffe. 2010. "Variational Inference for Large-Scale Models of Discrete Choice." *Journal of the American Statistical Association* 105:324–335.

Brick, J. Michael and Clyde Tucker. 2007. "Mitofsky-Waksberg: Learning from the Past." *Public Opinion Quarterly* 71:703–716.

Callegaro, Mario and Charles DiSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72:1008–1032.

Chaloner, Kathryn and Isabella Verdinelli. 1995. "Bayesian Experimental Design: A Review." *Statistical Science* 10:273–304.

Chen, Kuang, Harr Chen, Neil Conway, Joseph M. Hellerstein, and Tapan S. Parikh. 2010. "Usher: Improving Data Quality with Dynamic Forms." In *2010 IEEE 26th International Conference on Data Engineering (ICDE)*, pp. 321–332.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. 2007. "Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys." *Public Opinion Quarterly* 71:113–125.

Converse, Jean M. 1984. "Strong Arguments and Weak Evidence: The Open/Closed Questioning Controversy of the 1940s." *Public Opinion Quarterly* 48:267–282.

Converse, Jean M. 2009. *Survey Research in the United States: Roots and Emergence 1890-1960*. New Brunswick: Transaction Publishers.

Couper, Mick P. 2008. *Designing Effective Web Surveys*. Cambridge, UK: Cambridge University Press.

Couper, Mick P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75:889–908.

Couper, Mick P. and Peter V. Miller. 2009. "Web Survey Methods: Introduction." *Public Opinion Quarterly* 72:831–835.

Couper, Mick P., Roger Tourangeau, and Kristin Kenyon. 2004. "Picture This! Exploring Visual Effects in Web Surveys." *Public Opinion Quarterly* 68:255–266.

David, Herbert A. 1988. *The Method of Paired Comparisons*. London: Griffin, 2nd edition.

Dillman, Don A. 2002. "Presidential Address: Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early Twenty-First Century." *Public Opinion Quarterly* 66:473–494.

Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L. Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research* 38:1–18.

Dzyabura, Daria and John R. Hauser. 2011. "Active Machine Learning for Consideration Heuristics." *Marketing Science* (in press).

Farrell, Dan and James C. Petersen. 2010. "The Growth of Internet Research Methods and the Reluctant Sociologist." *Sociological Inquiry* 80:114–125.

Gal, David and Derek D Rucker. 2011. "Answering the Unasked Question: Response Substitution in Consumer Surveys." *Journal of Marketing Research* 48:185–195.

Gelfand, Alan E. and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85:398–409.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC, 2nd edition.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, Andrew and Kenneth Shirley. 2011. "Inference from Simulations and Monitoring Convergence." In *Handbook of Markov Chain Monte Carlo*, edited by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Boca Raton: CRC Press.

Gelman, Andrew, David A. van Dyk, Zaiying Huang, and John W. Boscardin. 2008. "Using Redundant Parameterizations to Fit Hierarchical Models." *Journal of Computational and Graphical Statistics* 17:95–122.

Glickman, Mark E. and Shane T. Jensen. 2005. "Adaptive Paired Comparison Design." *Journal of Statistical Planning and Inference* 127:279–293.

Groves, Robert M. 1990. "Theories and Methods of Telephone Surveys." *Annual Review of Sociology* 16:221–240.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75:861–871.

Groves, Robert M., Robert B. Cialdini, and Mick P. Couper. 1992. "Understanding The Decision to Participate in a Survey." *Public Opinion Quarterly* 56:475–495.

Groves, Robert M. and Steven G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169:439–457.

Hacker, Severin and Luis von Ahn. 2009. "Matchin: Eliciting User Preferences with an Online Game." *Proceedings of the 27th international conference on Human factors in computing systems* pp. 1207–1216.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*. Dordrecht: Springer, 2nd edition.

Holland, Jennifer L. and Leah Melani Christian. 2009. "The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys." *Social Science Computer Review* 27:196–212.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. New York: Wiley.

Kane, Michael J. and John W. Emerson. 2011. "bigmemory: Manage Massive Matrices With Shared Memory and Memory-Mapped Files." R package version 4.2.11.

Krause, Neal. 2002. "A Comprehensive Strategy for Developing Closed-Ended Survey Items for Use in Studies of Older Adults." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 57:S263–S274.

Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50:537–567.

Lazarsfeld, Paul F. 1944. "The Controversy Over Detailed Interviews – An Offer for Negotiation." *Public Opinion Quarterly* 8:38–60.

Lewry, Fraser and Tom Ryan. 2007. *Kittenwar: May the Cutest Kitten Win!* San Francisco: Chronicle Books.

Lindley, Dennis V. 1956. "On a Measure of the Information Provided by an Experiment." *Annals of Mathematical Statistics* 27:986–1005.

Lynch, Scott M. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer.

Merton, Robert K. 1987. "The Focussed Interview and Focus Groups: Continuities and Discontinuities." *Public Opinion Quarterly* 51:550–566.

Merton, Robert K. and Patricia L. Kendall. 1946. "The Focused Interview." *American Journal of Sociology* 51:541–557.

Mitofsky, Warren J. 1989. "Presidential Address – Methods and Standards: A Challenge for Change." *Public Opinion Quarterly* 53:446–453.

Mosteller, Frederick. 1951. "Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations." *Psychometrika* 16:3–9.

Newport, Frank. 2011. "Presidential Address: Taking AAPOR's Mission To Heart." *Public Opinion Quarterly* 75:593–604.

Presser, Stanley. 1990. "Measurement Issues in the Study of Social Change." *Social Forces* 68:856–868.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68:109–130.

R Development Core Team. 2011. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing: Version 2.13.1.

Resnick, Paul and Rahul Sami. 2008. "The Information Cost of Manipulation-Resistance in Recommender Systems." In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pp. 147–154.

Rossi, Peter E., Greg M. Allenby, and Rob McCulloch. 2006. *Bayesian Statistics and Marketing*. Hoboken, NJ: Wiley.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311:854–856.

Salganik, Matthew J. and Duncan J. Watts. 2009. "Web-based Experiments for the Study of Collective Social Dynamics in Cultural Markets." *Topics in Cognitive Science* 1:439–468.

Schuman, Howard. 2008. *Method and Meaning in Polls and Surveys*. Cambridge: Harvard University Press.

Schuman, Howard and Stanley Presser. 1979. "The Open and Closed Question." *American Sociological Review* 44:692–712.

Schuman, Howard and Jacqueline Scott. 1987. "Problems in the Use of Survey Questions to Measure Public Opinion." *Science* 236:957–959.

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

Shih, Tse-Hua and Xitao Fan. 2008. "Comparing Response Rates from Web and Mail Surveys: A meta-analysis." *Field Methods* 20:249–271.

Singh, Jagdip, Roy D. Howell, and Gary K. Rhoads. 1990. "Adaptive Designs for Likert-Type Data: An Approach for Implementing Marketing Surveys." *Journal of Marketing Research* 27:304–321.

Skitka, Linda J. and Edward G. Sargis. 2006. "The Internet as Psychological Laboratory." *Annual Review of Psychology* 57:529–555.

Smith, Tom W. 1987. "The Art of Asking Questions, 1936-1985." *Public Opinion Quarterly* 51:95–108.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Mallory Mcbride. 2009. "Open-ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly* 73:325–337.

Stern, Hal. 1990. "A Continuum of Paired Comparisons Models." *Biometrika* 77:265–273.

Stern, Hal. 1992. "Are All Linear Paired Comparison Models Empirically Equivalent?" *Mathematical Social Sciences* 23:103–117.

Sullivan, John L., James Piereson, and George E. Marcus. 1979. "An Alternative Conceptualization of Political Tolerance: Illusory Increases 1950s-1970s." *American Political Science Review* 73:781–794.

Thurstone, Leon L. 1927. "The Method of Paired Comparisons for Social Values." *Journal of Abnormal and Social Psychology* 21:384–400.

Toubia, Olivier and Laurent Flores. 2007. "Adaptive Idea Screening Using Consumers." *Marketing Science* 26:342–360.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press, 2nd edition.

Trautmann, Heike, Detlef Steuer, Olaf Mersmann, and Bjorn Bornkamp. 2011. "truncnorm: Truncated Normal Distribution." R package, version 1.0-5.

Urbanek, Simon. 2011. "multicore: Parallel Processing of R Code on Machines with Multiple Cores or CPUs." R package version 0.1-5.

van der Linden, Wim J. and Cees A. W. Glas. 2010. *Elements of Adaptive Testing*. New York: Springer.

von Ahn, Luis and Laura Dabbish. 2008. "Designing Games with a Purpose." *Communications of the ACM* 51:58–67.

Weinstein, Joshua R. 2009. *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.

Wickham, Hadley. 2011a. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40:1–29.

Wickham, Hadley. 2011b. "testthat: Get Started with Testing." *The R Journal* 3:5–10.

Wilkinson, Dennis M. 2008. "Strong Regularities in Online Peer Production." In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 302–309.

Yan, Ting, Frederick G. Conrad, Roger Tourangeau, and Mick P. Couper. 2011. "Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys." *International Journal of Public Opinion Research* 23:131–147.