# InterPoll: Crowd-Sourced Internet Polls

Benjamin Livshits and Todd Mytkowicz

Microsoft Research, US

## Abstract

Crowd-sourcing is increasingly being used to provide answers to online polls and surveys. However, existing systems, while taking care of the mechanics of attracting crowd workers, poll building, and payment, provide little to help the survey-maker or pollster in obtaining statistically significant results devoid of even the obvious selection biases.

This paper proposes InterPoll, a platform for programming of crowd-sourced polls. Pollsters express polls as embedded LINQ queries and the runtime correctly reasons about uncertainty in those polls, only polling as many people as required to meet statistical guarantees. To optimize the cost of polls, InterPoll performs query optimization, as well as bias correction and power analysis. The goal of InterPoll is to provide a system that can be reliably used for research into marketing, social and political science questions.

This paper highlights some of the existing challenges and how InterPoll is designed to address most of them. In this paper we summarize some of the work we have already done and give an outline for future work.

**1998 ACM Subject Classification** D.2.6 Programming Environments

**Keywords and phrases** CrowdSourcing, Polling, LINQ

**Digital Object Identifier** 10.4230/LIPIcs.SNAPL.2015.156

## 1 Introduction

Online surveys have emerged as a powerful force for assessing properties of the general population, ranging from marketing studies, to product development, to political polls, to customer satisfaction surveys, to medical questionnaires. Online polls are widely recognized as an affordable alternative to in-person surveys, telephone polls, or face-to-face interviews. Psychologists have argued that online surveys are far superior to the traditional approach of finding subjects which involves recruiting college students, leading to the famous quip about psychology being the study of the college sophomore [16].

Online surveys allow one to reach wider audience groups and to get people to answer questions that they may not be comfortable responding to in a face-to-face setting. While online survey tools such as Instant.ly, SurveyMonkey, Qualtrics, and Google Customer Surveys take care of the mechanics of online polling and make it easy to get *started*, the results they produce often create more questions than they provide answers [18, 25, 21, 39, 106, 48].

Surveys, both online and offline, suffer from *selection biases*, as well as non-response, and coverage issues. These biases are not trivial to correct for, yet without doing so, the data obtained from surveys may be less than representative, which complicates generalizing to a larger population. InterPoll allows the developer to both *estimate* and *correct* for the biases and errors inherent in the data they are collecting.

It is also not so obvious how many people to poll. Indeed, polling too few yields results that are not statistically significant; polling too many is a waste of money. None of the current survey platforms help the survey-maker with deciding on the appropriate number of samples. Today's online survey situation can perhaps be likened to playing slot machines

1st Summit on Advances in Programming Languages (SNAPL'15).
Eds.: Thomas Ball, Rastislav Bodík, Shriram Krishnamurthi, Benjamin S. Lerner, and Greg Morrisett; pp. 156–176
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

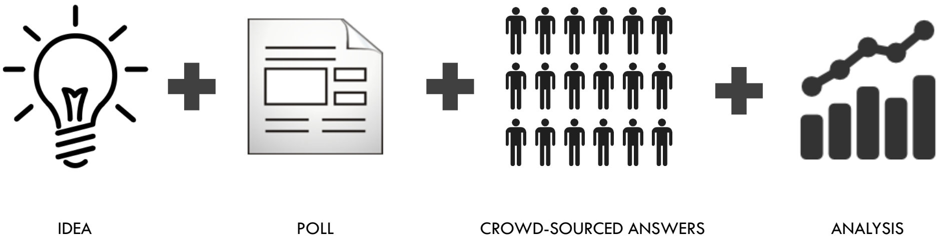
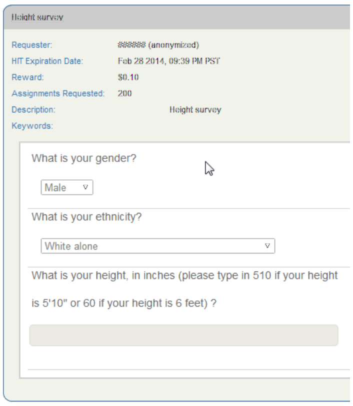


**Figure 1** Conceptual architecture of InterPoll highlighting the major steps.

with today's survey sites playing the role of a casino; it is clearly in the interest of these survey sites to encourage more polls being completed.

In addition to the issue of data quality and representativeness, *cost* of the polls is an important consideration for poll makers, especially given that thousands of participants may be required. Even if answering a single question can costs cents, often getting a high level of assurance for targeted population segment involves hundreds of survey takers at significant cost. In fact, deciding on how to properly target the survey is a non-trivial task: if general audience surveys cost $.10 per question and *targeted* ones cost $.50 per question, is it better to ask five times as many questions of the general audience and then post-process the results or is it better to ask fewer questions of the targeted audience? Given that demographic targeting can often involve dozens of categories (males, 20–30, employed full-time, females, 50–60, employed part-time, females, 20–30, students, etc.) how does one properly balance the need for targeted answers and the cost of reaching these audiences?



Height survey

Requester: ###### (anonymized)
HIT Expiration Date: Feb 28 2014, 09:39 PM PST
Reward: $0.10
Assignments Requested: 200
Description: Height survey
Keywords:

What is your gender?
Male

What is your ethnicity?
White alone

What is your height, in inches (please type in 510 if your height is 5'10" or 60 if your height is 6 feet) ?

**Figure 2** Sample form produced by InterPoll for the Mechanical Turk back-end.

We see these challenges as interesting *optimization problems*. To address some of these issues, InterPoll has an optimization engine whose goals is to determine (a sequence of) questions to ask and targeting restrictions to use. The primary goal of the optimization is to get a certain level of certainty in a developer-provided question (i.e. do men aged between 30–50 prefer *Purina Dog Chow* to *Precise Naturals Grain Free*), while minimizing the cost involved in running the poll on a large scale.

**This Paper:** This paper presents a high-level summary of InterPoll, a platform for in-application scripting of crowd-sourced polls, giving developers streamlined access to crowd-sourced poll data. The processing pipeline of InterPoll is shown in Figure 1. In this paper we briefly summarize the research on InterPoll we have already done and outline some of the avenues for future work.

InterPoll is an attempt to balance human and machine computation. One of the goals is to allow for easy integration with existing programs. As a result, we opted to use LINQ queries [65] already widely used by developers, instead of proposing a DSL. InterPoll is implemented as a library that can be linked into an existing application.

Note that due to LINQ's runtime introspection features, this allows us to effectively build an optimizing runtime for surveys within a library. InterPoll performs query optimization [59], as well as bias correction and power analysis [60], among other features, to enable a system that can be reliably used for research in marketing, social, and political sciences.

### Motivating Examples

As mentioned above, one of the goal of InterPoll is to make running crowd-sourced polls easy for the developer. We accomplish this by using LINQ [65], language-integrated queries. LINQ is natively supported by .NET, with Java providing similar facilities with JQL.

▶ **Example 1** (Basic filtering)**.** A simple poll may be performed the following way.

```
1  var people = new MTurkQueryable<Person>(true, 5, 100, 2);
2  var liberalArtsPairs = from person in people
3           where person.Employment == Employment.STUDENT
4           select new {
5                  Person = person,
6                  Value = person.PoseQuestion<bool>(
7                         "Are you a liberal arts major?")
8                  };
```

The first line gets a handle to a population of users, in this case obtained from Mechanical Turk, although other back-ends are also possible. Populations on which we operate have associated demographic information; for example, note that the `where` clause on line 3 ensures that we only query (college) students.

This poll will ask (college) students if they study liberal arts, producing an iterator of ⟨`Student`, `bool`⟩ pairs represented in .NET as `IEnumerable`.

▶ **Example 2** (Counting)**.** Given `liberalArtsPairs`,

```
var libralArtMajors = from pair in liberalArtsPairs where pair.Value == true select person;
double percentage = 100.0 * libralArtMajors.Count() / liberalArtsPairs.Count();
```

it is possible to do a subsequent operation on the result, such as printing out all pairs or using, the `Count` operation to count the liberal arts majors. The last line computes the percentage of liberal art majors within the previously collected population.

▶ **Example 3** (Uncertainty)**.** InterPoll explicitly supports computing with uncertain data, using a style of programming proposed in Bornholt *et al.* [10].

```
1  var liberalArtWomen = from person in people where person.Gender == Gender.FEMALE
2    where person.Employment == Employment.STUDENT select person.PoseQuestion<bool>("Are you a liberal arts major?");
3
4  var liberalArtMen = from person in people where person.Gender == Gender.MALE
5    where person.Employment == Employment.STUDENT select person.PoseQuestion<bool>("Are you a liberal arts major?");
6
7  var femaleVar = liberalArtWomen.ToRandomVariable(); var maleVar = liberalArtMen.ToRandomVariable();
8  if (femaleVar > maleVar) Console.WriteLine("More female liberal arts majors.");
9  else Console.WriteLine("More male liberal arts majors.");
```

Here, we convert the Boolean output of the posted question to a random variable (line 7). Then we proceed to compare these on line 8. Comparing two random variables (`femaleVar` and `maleVar`) results in a Bernoulli which the C# type system then implicitly casts (i.e., in the > comparison on line 8) into a boolean by running a t-test on the resulting Bernoulli[10].

▶ **Example 4** (Explicit t-tests)**.** Here we explicitly perform the t-test at a specified confidence interval.

```
var test = liberalArtMen.ToRandomVariable() > liberalArtWomen.ToRandomVariable();
if (test.AtConfidence(.95)) { ... }
```

The test and the confidence interval determine the outcome of a power analysis that InterPoll will perform to decide how many (male and female) subjects to poll.

▶ **Example 5** (Optimizations). Suppose we are conducting a marketing study of dog owners' preference for Purina Puppy Chow. Specifically, we are trying decide if married women's attitude toward this product is more positive than that of married men.

```
var puppyChowWomen = from person in people where person.PoseQuestion<bool>("Are you a dog owner?")
          == true where person.Gender == Gender.FEMALE where person.Relationship == Relationship.MARRIED
   select person.PoseQuestion<bool>("Would you consider using Purina Puppy Chow?");
```

Similarly, for men:

```
var puppyChowMen = from person in people where person.PoseQuestion<bool>("Are you a dog owner?"))
          == true where person.Gender == Gender.MEN where person.Relationship == Relationship.MARRIED
   select person.PoseQuestion<bool>("Would you consider using Purina Puppy Chow?");
```

To compare these two, the following comparison may be used:

```
if (puppyChowWomen > puppyChowMen) Console.WriteLine("Women like puppy chow more");
```

In this case it is not so obvious how to sample from the population: a naïve strategy is to sample women first, then sample men. However, another strategy may be to sample everyone (who is `MARRIED`) and to separate them into two streams: one for women, the other for men. Lastly, sampling from the same population is likely to yield a disproportional number of samples in either population. For example, 64% of users of the uSamp platform are women [101] as opposed to 51%, as reported by the US 2012 Census. InterPoll's LINQ abstractions let a polster specify what to query and leaves the particular strategy for implementing a poll to InterPoll's optimizations.

## Challenges

The examples described above raise a number of non-trivial challenges.

**Query optimization:** How should these queries be executed? How can the queries be optimized to avoid unnecessary work? Should doing so take the surrounding .NET code into which the queries are embedded into account? Should they be run independently or should there be a degree of reuse (or result caching) between the execution plans for the men and women? While a great deal of work on database optimizations exist, both for regular and crowd-sourced databases, much is not directly applicable to the InterPoll setting [14, 38, 69, 83, 9, 68], in that the primary goal of InterPoll optimizations is reducing the amount of money spent on a query.

**Query planning:** How should we run a given query on the crowd back-end? For instance, should pre-filter crowd workers or should we do post-filtering ourselves? Which option is cheaper? Which crowd back-end should we use, if they have different pricing policies? Should the filtering (by gender and relationship status) take place as part of population filtering done by the crowd provider?

**Bias correction:** Given that men and women do not participate in crowd-sourcing at the same rate (on some crowd-sourcing sites, one finds about 70% women and 30% men [78, 42, 71]), how do we correct for the inherent population bias to match the more equal gender distribution consistent with the US Census? Similarly, studies of CrowdFlower samples show a disproportionately high number of democrats vs. republicans [27]. Mobile crowd-sourcing attracts a higher percentage of younger participants [32].

**Ignorable sample design:** Ignorable designs assume that sample elements are missing from the sample when the mechanism that creates the missing data occurs at random, often referred to as missing at random or completely missing at random [77]. An example of non-ignorable design is asking what percentage of people know how to use a keyboard:

in a crowd sample that need a keyboard to fill out the survey, the answer is likely to be nearly 100%; in the population as a whole it is likely to be lower.

**Power analysis:** Today, the users of crowd-sourcing are forced to decide how many participants or workers to use for every task, yet there is often no solid basis for such a decision: too few workers will produce results of no statistical significance; too many will result in over-payment. How many samples (or workers) are required to achieve the desired level of statistical significance?

**Crowd back-end selection:** Given that different crowd back-ends may present different cost trade-offs (samples stratified by age or income may be quite costly, for example) and demographic characteristics, how do we pick an optimal crowd for running a given set of queries [67]? How do we compare query costs across the back-ends to make a globally optimal decision?

**Quality control:** What if the users are selecting answers at random? This is especially an issue if we ask about properties that eschew independent verification without direct contact with the workers, such as their height. A possible strategy is to insert attention-checking questions also called "catch trials" and the like [70, 46, 96].

**Privacy of human subjects:** Beyond the considerations of ethics review boards (IRBs) and HIPAA rules for health-related polls, there is a persistent concern about being able to de-anonymize users based on their partial demographic and other information.

### Domains

InterPoll lowers the effort and expertise required for non-expert programmers to specify polls which benefits many non-experts in some of the following domains.

**Social sciences:** social sciences typically rely on data obtained via studies and producing such data is often difficult, time-consuming, and costly [25]. While not a panacea, online polls provide a number of distinct advantages [53].

**Political polls:** these are costly and require a fairly large sample size to be considered reliable. By their very nature, subjects from different geographic locales are often needed, which means that either interviewers need to cast a wide net (exit polling in a large number of districts) [88, 8] or they need to conduct a large remote survey (such as telephone surveys) [107, 30, 90].

**Marketing polls:** While much has been written about the upsides and downsides of online surveys [25, 77, 2, 22], the ability to get results cheaply, combined with the ease of targeting different population segments (i.e., married, high income, dog owners) makes the web a fertile ground for marketing polls. Indeed, these are among the primary uses of sites such as Instant.ly [101], Survey Monkey [41], and Google Customer Surveys [66].

**Health surveys:** A lot of researchers have explored the use of online surveys for collecting health data [93, 76, 26, 94, 5].

In all of the cases above, in addition to population biases, so-called *mode effects*, i.e. differences in results caused by asking questions online vs. on the telephone vs. in person are possible [12, 87, 109, 23, 32, 82, 80, 107, 30, 90, 47, 6].

## 2 Three Themes

In this section, we cover three themes we have explored in our research focusing on InterPoll so far. Section 2.1 talks about optimizing InterPoll queries. Section 2.2 discusses power analysis. Lastly, Section 2.3 gives an overview of unbiasing.

```
from structure in from person in employees where person.Wage > 4000 && person.Region == "WA"
    select new { Name = person.Name, Boss = person.GetBoss(), Sales = person.GetSales(42) }
where structure.Sales.Count > 5 select structure.Boss;
```

```
from person in employees where (person.Wage > 4000 && person.Region == "WA") && (person.GetSales(42).Count > 5)
 select person.GetBoss();
```

**Figure 3** Query flattening.

## 2.1 Optimizations

We have applied two kinds of optimizations to InterPoll queries: static and runtime optimizations [59]. The former can be used to address poor programming practices used by developers or users who program InterPoll queries. Additionally, static optimizations are used to *normalize* InterPoll queries before runtime optimizations can be applied to them. In optimizing queries in InterPoll, we try to satisfy the following goals:

- Reduce **overall cost** for running a query; clearly for many people, reducing the cost of running survey is the most important "selling feature" when it comes to optimizations. Not only does it allow people with a low budget to start running crowd-sourced queries, it also allows survey makers to 1) request more samples and 2) run their surveys more frequently. Consider someone who may previously have been able to run surveys *weekly* now able to do so *daily*.
- Reduce the **end-to-end time** for running a query; we have observed that in many cases, making surveys requires *iterating* on how the survey is formulated. Clearly, reducing the running times allows the survey maker to iterate over their surveys to refine the questions much faster. Consider someone who needs to wait for week only to discover that they need to reformulate their questions and run them again.
- Reduce the **error rate** (or confidence interval) for the query results; while we support unbiasing the results in InterPoll, one of the drawbacks that is often cited as a downside of unbiasing is that the *error rate* goes up. This is only natural: if we have an unrepresentative sample which we are using for extrapolating the behavior for the overall population, the high error rate will capture the paucity of data we are basing our inference on.

LINQ provides an accessible mechanism to access the internal parts of a LINQ query via LINQ Expressions. Each LINQ query is translated into an expression AST, which can then be traversed and rewritten by LINQ Providers. InterPoll provides an appropriate set of visitors that rewrite LINQ query trees so as to both optimize them and also connect those query trees to the actual Mechanical Turk crowd. This latter "plumbing" is responsible for obtaining Mechanical Turk data in XML format and then at runtime parsing and validating it, and embedding the data it into type-safe runtime data structures.

### 2.1.1 Static Optimizations

Figure 3 gives an example of *flattening* LINQ expression trees and eliminating intermediate structures that may otherwise be created. In this example, `structure` is "inlined" into the query so that all operations are on the `person` value.

The other two optimizations we have implemented are *query splitting* which separates general and demographic questions that filter based on demographic characteristics into a where `clause` and *common subexpressions elimination* which identifies and merges common subexpressions in LINQ queries.

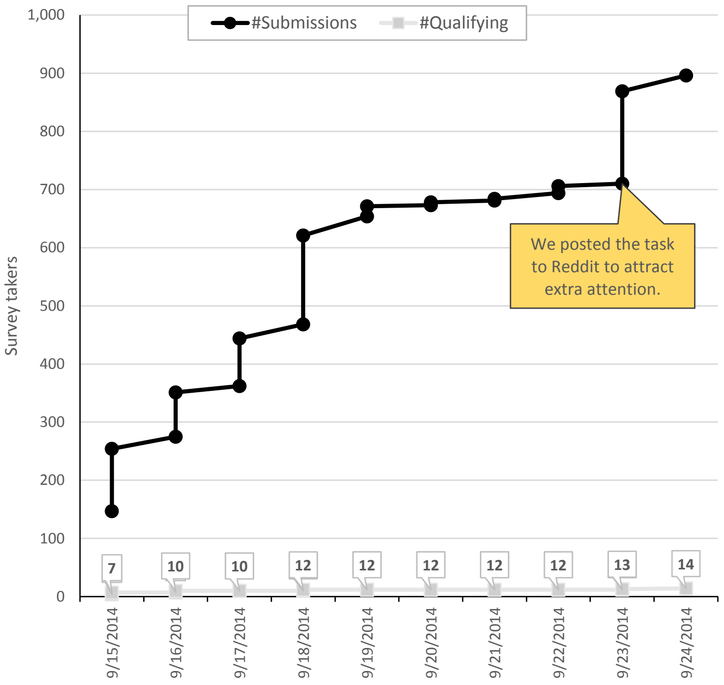


**Figure 4** Completions for qualifications and passing for the budget query in Figure 5.

### 2.1.2 Dynamic Optimizations

However, runtime optimizations are generally more fruitful and, as mentioned above, can be implemented as a LINQ rewriting step that at runtime can change the evaluation strategy in response to statistics that are gathered at runtime.

**Yield:** The *yield optimization* addresses the problem of low-yield: this is when `where` clauses return only a very small percentage of the sampled participants, yet we have to pay for all of them. We have explored several strategies for evaluating such queries and evaluated the trade-offs between the cost of obtaining a certain number of responses and the completion time. Note that in practice, it is not uncommon to wait a week (or more) for certain uncommon demographics.

Figure 4 shows a long-running query where we ask for people's attitudes toward the US Federal budge deficit, with the where clause being:

```
person.Income==INCOME_35_000_TO_49_999 &&
person.Ethnicity==BLACK_OR_AFRICAN_AMERICAN
```

The query is shown in Figure 5.

The graph shows that despite about 900 people taking this poll over a full week only 14 meet the demographic constraints. By using *qualification tests*, an MECHANICAL TURK mechanism which requires a user successfully answer "qualification questions" (at low cost) before being able to complete the full survey (at full cost), INTERPOLL's optimized evaluation strategy can reduce the overall cost of this query by 8×.

**Rebalancing:** INTERPOLL supports answering decision questions of the form $r_1$ `boolOp` $r_1$, where both $r_1$ and $r_2$ are random variables obtained from segments of the population. To

```
var query = from person in people select new {
     Attitude = person.PoseQuestion(
          "How do you think the US Federal Government's yearly budget deficit has changed since January 2013?",
          "Increased a lot ", "Increased a little ", "Stayed about the same", "Decreased a little ", "Decreased a lot "),
     Gender = person.Gender, Income = person.Income, Ethnicity = person.Ethnicity ,
};
query = from person in query where person.Income == Income.INCOME_35_000_TO_49_999
                    && person.Ethnicity == Ethnicity.BLACK_OR_AFRICAN_AMERICAN select person;
```

**Figure 5** Budget deficit attitudes query.

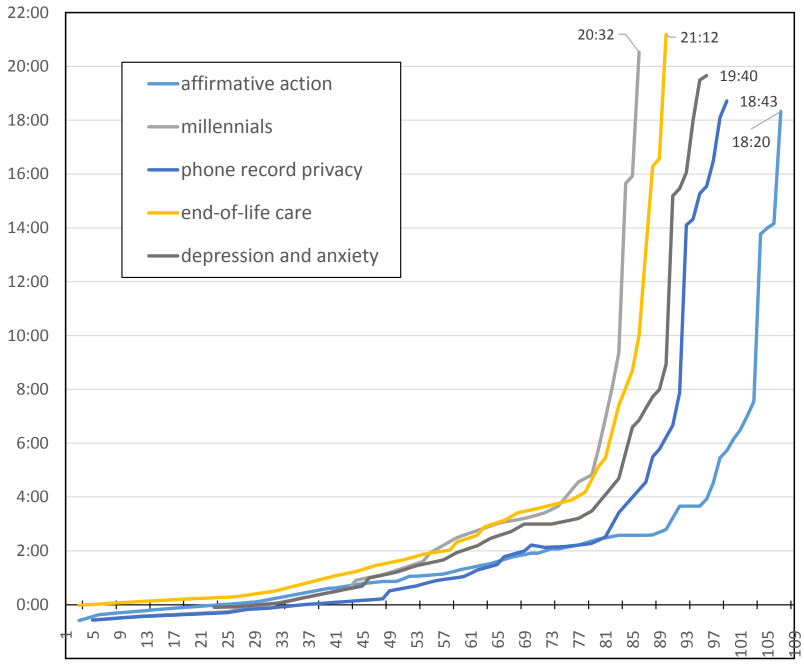


**Figure 6** Time savings attained through the rebalancing optimization.

answer such *decision queries*, InterPoll repeatedly considers *pairs* of people from the categories on the left and right hand sides and then performs a sequential probability ratio test [60] to decide how many samples to request. A common problem, however, is that the two branches of the comparison are *unbalanced*: we are likely to have an unequal number of males and females in our samples, or people who are rich or poor, or people who own and do not own dogs.

The main mechanism for *rebalancing optimizations* is to reward the sub-population that is scarce more and the one that is plentiful less by adjusting the payoff of the Mechanical Turk task. Figure 6 shows the effect of increasing the reward by measuring the time difference to get to $x$ completes between the default strategy and the fast branch of our rebalancing strategy. While the number of completes (people) is shown on the $x$ axis, the times are measured in hours, shown on the $y$ axis. Overall, the time savings achieved through rebalancing are significant: the fast branch gets to 70 completes over 2 hours faster and, for all strategies, to get to 90 completes, the it takes up to 21 more hours.

**Panel building:** The last optimization is motivated by the desire to avoid unbiasing by constructing *representative* population samples. Unbiasing based on unrepresentative samples frequently widens the confidence interval and is, therefore, less than desirable. Just like in real world polling, in InterPoll we have created ways to pre-build representative panels. An illustrative example is unbiasing height values (in cm) based on *ethnicity*. We limit our analysis to 50 respondents. Out of those, 35 were male and 15 female. Considering the female sub-population only, the mean height is *virtually the same* as the Wikipedia value of height in the US. Unbiasing the female heights with respect to ethnicity, however, produces 166.15 ± 5.32 cm. This is significantly larger than the true value and the difference emerges when we discover that our sample is reweighed using a weight of 50 for African Americans (12% in the population but only 1 in our sample). This taller woman (188 cm) has a large effect on the unbiased mean.

## 2.2 Power Analysis

We have implemented support for power analysis using an approach referred to as sequential acceptance sampling (SPRT) [102]. This approach to hypothesis testing requires drawing (pairs of) samples from Mechanical Turk until a convergence condition is met. The convergence condition is calculated based on the number of positive and negative responses. While the details of this process are discussed in a prior paper [60], below we show some queries and the number of samples required to resolve each. These queries are taken from ten real-life debates conducted via the Intelligence Squared site `http://intelligencesquaredus.org`.

Figure 7 shows a summary of our results for each of the debate polls. Alongside the outcome of each poll, we show the power analysis-computed power. We also show the dollar cost required to obtain the requisite number of samples from the crowd. `ObesityIsGovernmentBusiness` was the most costly debate of them all, requiring 265 workers, of which 120 (45%) were yes votes, whereas 145 (55%) said no.

| Task | Outcome | Power | Cost |
|---|---|---|---|
| `MilennialsDontStandAChance` | No | 37 | $3.70 |
| `MinimumWage` | No | 43 | $4.30 |
| `RichAreTaxedEnough` | No | 51 | $5.10 |
| `EndOfLife` | No | 53 | $5.30 |
| `BreakUpTheBigBanks` | Yes | 73 | $7.30 |
| `StrongDollar` | No | 85 | $8.50 |
| `MarginalPower` | No | 89 | $8.90 |
| `GeneticallyEngineeredBabies` | Yes | 135 | $13.50 |
| `AffirmativeActionOnCampus` | Yes | 243 | $24.30 |
| `ObesityIsGovernmentBusiness` | No | 265 | $26.50 |

**Figure 7** Ten debates: outcomes, power analysis, costs.

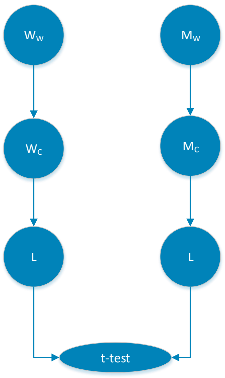
## 2.3 Unbiasing

During the 1936 U.S. presidential campaign, the popular magazine Literary Digest conducted a mail-in election poll that attracted over two million responses, a huge sample even by today's standards. Literary Digest notoriously and erroneously predicted a landslide victory for Republican candidate Alf Landon. In reality, the incumbent Franklin D. Roosevelt decisively won the election, with a whopping 98.5% of the electoral vote, carrying every state except for Maine and Vermont. So what went wrong? As has since been pointed out, the magazine's forecast was based on a highly non-representative sample of the electorate – mostly car and telephone owners, as well as the magazine's own subscribers – which underrepresented Roosevelt's core constituencies. By contrast, pioneering pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but representative samples to predict the election outcome with reasonable accuracy. This triumph of brains over brawn effectively marked the end of convenience sampling, and ushered in the age of modern election polling.

It is broadly acknowledged that while crowd-sourcing platforms present a number of exciting new benefits, conclusions that may result from crowd-sourced experiments need to be treated with care [7, 6]. External validity is an assessment of whether the causal estimates deduced from experimental research would persist in other settings and with other samples. For instance, concerns about the external validity of research conducted using student samples (the so-called college sophomore problem) have been debated extensively [16].

The composition of population samples found on crowd-sourcing sites such as Mechanical Turk generally differs markedly from the overall population, leading some researchers to question the overall value of online surveys [6, 35, 25, 53, 78, 42, 71, 43, 11, 3, 52, 81, 10]. Wauthier *et al.* [103] advocate a bias correction approach for crowd-sourced data that we generally follow.

▶ **Example 6** (Unbiasing)**.** Consider deciding if there are more female liberal art majors than than the there are male ones. The ultimate comparison will be performed via a t-test.
However, the first task is to determine the expected value of female and male liberal art majors given that we drew $S$ samples from the crowd. These values can be computed as shown below: $E[L_W|C] = Pr[L|W_C] \times Pr[W_C|W_W] \times S$ and $E[L_M|C] = Pr[L|M_C] \times Pr[M_C|M_W] \times S$ where $L_W$ and $L_M$ are the number of female and male liberal art major, respectively, $W_C$ and $M_C$ stand for a woman/man being in the crowd, and $W_W$ and $M_W$ stand for a woman/man being the world as reflected by a broad population survey such as the US census; the latter two probabilities may be related at 51 : 49, for example.
Note that our goal is to discern the expected value of liberal art majors per gender *in the world*. We can unbias our data by using the probability of observing a woman in the crowd given there is a woman in the world: $E[W_L|W] = E[W_L|C] \times P(W_C|W_W)$ and similarly for men $E[M_L|M] = E[M_L|C] \times P(M_C|M_W)$.



While $E[W_L|C]$ and $E[M_L|C]$ can be approximated by observing the crowd-sourced results for the female and male sub-segments of the population, coefficients such as $P(W_C|W_W)$ can be computed from our knowledge of crowd population vs. that in the world in general. For example, if women to men are at 50%:50% in the world and at 30%:70% in the crowd, $P(W_C|W_W) = .7$ and $P(M_C|M_W) = .3$.

Note that the above example presents a simple model that does not, for example, explicitly represent the factor of ignorability [33], pg. 202 of our experimental design. Also note that unbiasing generally may need to be done before we perform a t-test to reshape the underlying distributions.

## 3 Related Work

There are several bodies of related work from fields that are usually not considered to be particularly related, as outlined below.

### 3.1 Crowd-Sourcing Systems

There has been a great deal of interest in recent years in building new systems for automating crowd-sourcing tasks.

**Toolkits:** TurKit [58] is one of the first attempts to automate programming crowd-sourced systems. Much of the focus of TurkIt is the iterative paradigm, where solutions to crowd-sourced tasks are refined and improved by multiple workers sequentially. The developer can write TurkIt scripts using JavaScript. AutoMan [4] is a programmability approach to combining crowd-based and regular programming tasks, a goal shared with Truong *et al.* [98]. The focus of AutoMan is on computation reliability, consistency and accuracy of obtained results, as well as task scheduling. Turkomatic [57, 56] is a system for expression crowd-sourced tasks and designing workflows. CrowdForge is a general purpose framework for accomplishing complex and interdependent tasks using micro-task markets [54]. Some of the tasks involve article writing, decision making, and science journalism, which demonstrates the benefits and limitations of the chosen approach. More recently, oDesk has emerged as a popular marketplace for skilled labor. CrowdWeaver is a system to visually manage complex crowd work [51]. The system supports the creation and reuse of crowd-sourcing and computational tasks into integrated task flows, manages the flow of data between tasks, etc.

Wiki surveys [79] is a novel approach of combining surveys and free-form interviews to come up to answers to tough questions. These answers emerge as a result of pair-wise comparisons of individual ideas volunteered by participants. As an example, 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, with subsequent data analysis employed to estimate "public opinion" based on a large number of pair-wise outcomes.

We do not aim to adequately survey the vast quantity of crowd-sourcing-related research out there; the interested reader may consult [108]. Notably, a great deal of work has focused on matching users with tasks, quality control, decreasing the task latency, etc.

Moreover, we should note that our focus is on *opinion polls* which distinguishes InterPoll work from the majority of crowd-sourcing research which generally requires giving solutions to a particular task, such as deciphering a license plate number in a picture, translating sentences, etc. In InterPoll, we are primarily interested in self-reported opinions of users about themselves, their preferences, and the world at large.

**Some important verticals:** Some crowd-sourcing systems choose to focus on specific verticals. The majority of literature focuses on the following four verticals:

- social sciences [27, 5, 3, 53, 13, 11, 16, 35, 74];
- political science and election polls [90, 6, 7, 87, 5, 47, 107];
- marketing [41, 101, 25]; and
- health and well-being [93, 94, 26, 76, 106, 5, 7, 2, 81, 19].

## 3.2 Optimizing Crowd Queries

CrowdDB [29] uses human input via crowd-sourcing to process queries that regular database systems cannot adequately answer. For example, when information for `IBM` is missing in the underlying database, crowd workers can quickly look it up and return as part of query results, as requested. CrowdDB uses SQL both as a language for posing complex queries and as a way to model data. While CrowdDB leverages many aspects of traditional database systems, there are also important differences. CrowdDB extends a traditional query engine with a small number of operators that solicit human input by generating and submitting work requests to a microtask crowd-sourcing platform. It allows any column and any table to be marked with the `CROWD` keyword. From an implementation perspective, human-oriented query operators are needed to solicit, integrate and cleanse crowd-sourced data. Supported crowd operators include `probe`, `join`, and `compare`.

Marcus *et al.* [61, 62, 63] have published a series of papers outlining a vision for Qurk, a crowd-based query system for managing crowd workflows. Some of the motivating examples [61] include identifying people in photographs, data discovery and cleansing (who is the CEO of a particular company?), sentiment identification in Twitter messages, etc.

Qurk implements a number of optimizations [63], including task batching, replacing pairwise comparisons with numerical ratings, and pre-filtering tables before joining them, which dramatically reduces the overall cost of sorts and joins on the crowd. End-to-end experiments show cost reductions of $14.5x$ on tasks that involve matching up photographs and ordering geometric pictures. These optimization gains in part inspire our focus on cost-oriented optimizations in InterPoll.

Marcus *et al.* [62] study how to estimate the *selectivity* of a predicate with help from the crowd, such as filters photos of people to those of males with red hair. Crowd workers are shown pictures of people and provide either the gender or hair color they see. Suppose we could estimate that red hair is prevalent in only 2% of the photos, and that males constitute 50% of the photos. We could order the tasks to ask about red hair first and perform fewer HITs overall. Whereas traditional selectivity estimation saves database users time, optimizing operator ordering can save users money by reducing the number of HITs. We consider these estimation techniques very much applicable to the setting of InterPoll, especially when it comes to free-form `PoseQuestion`, where we have no priors informing us of the selectivity factor of such a filter. We also envision of a more dynamic way to unfold questions in an order optimized for cost reduction.

Kittur *et al.* [51] present a system called CrowdWeaver, designed for visually creating crowd workflows. CrowdWeaver system supports the creation and reuse of crowd-sourcing and computational tasks into integrated task flows, manages the flow of data between tasks, and allows tracking and notification of task progress. While our focus in InterPoll is on embedding polls into general-purpose programming languages such as C#, InterPoll could definitely benefit from a visual task builder approach, so we consider CrowdWeaver complimentary.

Somewhat further afield, Gordon *et al.* [34] describe a language for probabilistic programming and give an overview of related work. Nilesh *et al.* [20] talk about *probabilistic databases* designed to work with imprecise data such as measured GPS coordinates, and the like.

## 3.3 Database and LINQ Optimizations

While language-integrated queries are wonderful for bringing the power of data access to ordinary developers, LINQ queries frequently do not result in most efficient executions. There has also been interest in both formalizing the semantics of [14] and optimizing LINQ queries.

Grust *et al.* propose a technique for alternative efficient LINQ-to-SQL:1999 compilation [38]. Steno [68] proposes a strategy for removing some of the inefficiency in built-in LINQ compilation and eliminates it by fusing queries and iterators together and directly compiling LINQ queries to .NET code.

Nerella *et al.* [69] relies on programmer-provided annotations to devise better queries plans for language-integrated queries in JQL, Java Query Language. Annotations can provide information about shapes of distribution for continuous data, for example. Schueller *et al.* [83] focus on bringing the idea of *update propagation* to LINQ queries and combining it with reactive programming. Tawalare *et al.* [95] explore another compile-time optimization approach for JQL.

Bleja *et al.* [9] propose a new static optimization method for object-oriented queries dealing with a special class of sub-queries of a given query called "weakly dependent sub-queries." The dependency is considered in the context of SBQL non-algebraic query operators like *selection* and *projection*. This research follows the stack-based approach to query languages.

## 3.4 Web-Based Polls and Surveys

Since the time the web has become commonplace for large segments of the population, we have seen an explosion of interest in using it as a means for conducting surveys. Below we highlight several papers in the growing literature on this subject [2, 5, 17, 18, 21, 22, 25, 28, 44, 31, 35, 36, 37, 39, 45, 49, 50, 52, 53, 64, 81, 84, 89, 97, 106, 2, 5, 17, 18, 21, 22, 25, 28, 44, 31, 35, 36, 37, 39, 45, 49, 50, 52, 53, 64, 81, 84, 89, 97, 106].

**Online Demographics:** Recent studies reveal much about the demographics of crowd-sourcing sites such as Amazon's Mechanical Turk [6, 35, 25, 53, 104, 23, 78, 42, 71, 43, 11, 3, 52, 81, 74]. Berinsky *et al.* [6] investigate the characteristics of samples drawn from the Mechanical Turk population and show that respondents recruited in this manner are often *more* representative of the U.S. population than in-person convenience samples – the modal sample in published experimental political science – but *less* representative than subjects in Internet-based panels or national probability samples. They succeeded in replicating three experiments, the first one of which focuses on welfare spendings or assistance to the poor. They compared Mechanical Turk results with those obtained via the General Social Surveys (GSS), a nationally-representative face-to-face interview sample. While subtle differences exist, the overall results were quite similar between the GSS and Mechanical Turk (37% vs 38%). The second experiment involves replicating the so-called *Asian disease* experiment, which involves asking respondents to choose between two policy options. The results were comparable to those obtained in the original experiment by Tversky and Kahneman [99] on a student sample. The last experiment is described in Kam *et al.* [85] and involves measuring the preference for a risky policy option over a certain policy option. Additionally, Berinsky *et al.* discuss the internal and external validity threats. These three experiments provide a diverse set of studies to reproduce using InterPoll.

Ipeirotis [43, 42] focuses his analysis on the *demographics* of the Mechanical Turk marketplace. Overall, they find that approximately 50% of the workers come from the United States and 40% come from India. Significantly more workers from India participate on Mechanical Turk because the online marketplace is a primary source of income, while in the US most workers consider Mechanical Turk a secondary source of income. While money is a primary motivating reason for workers to participate in the marketplace, workers also cite a variety of other motivating reasons, including entertainment and education. Along with other studies, Ipeirotis provides demographic comparisons for common categories such as gender, age, education level, household income, and marital status for both countries. Ipeirotis [42] digs deeper into worker motivation, cost vs. the number of workers interested, time of completion vs. reward, etc. We believe that this data can be useful to give more fine-grained cost predictions for InterPoll queries and producing more sophisticated query plans involving tasks priced at various levels, for example. Additionally, while our initial focus is on query cost, we should be able to model completion rates fairly precisely as well. Of course, demographic data is also important for unbiasing query results.

Paolacci *et al.* [71] compare different recruiting methods (lab, traditional web study, web study with a specialized web site, Mechanical Turk) and discuss the various threats to validity. They also present comparisons of Mechanical Turk samples with those found through subject recruitment at a Midwestern university and through several online discussion boards that host online experiments in psychology, revealing drastic differences in terms of the gender breakdown, average age, and subjective numeracy. The percentage of failed catch trials varied as well, but not drastically; Mechanical Turk workers were quite motivated to complete the surveys, compared to those found though online discussion boards. While data quality does not seem to be adversely affected by the task payoff, researcher reputation might suffer as a result of poor worker perception and careless researchers "black-listed" on sites such as `http://turkopticon.differenceengines.com`.

Ross *et al.* [78] describe how the worker population has changed over time, shifting from a primarily moderate-income, U.S.-based workforce toward an increasingly international group, with a significant population of young, well-educated Indian workers. This change in population points to how workers may treat Turking as a full-time job, which they rely on to make ends meet. The paper contains comparisons across nationality, gender, age, and income, pinpointing a trend toward a growing number of young, male, Indian Turkers. Interesting opportunities exist for cost optimizations in InterPoll if we determine that different worker markets can provide comparable results (for a given query), yet are priced differently.

Buhrmester *et al.* [11] report that demographic characteristics suggest that Mechanical Turk participants are at least as diverse and more representative of non-college populations than those of typical Internet and traditional samples. Most importantly, they found that the quality of data provided by Mechanical Turk met or exceeded the psychometric standards associated with published research.

Andreson *et al.* [1] report that Craigslist can be useful in recruiting women and low-income and young populations, which are often underrepresented in surveys, and in recruiting a racially representative sample. This may be of particular interest in addressing recruitment issues in health research and for recruiting non-WEIRD (Western, Educated, Industrialized, Rich, Democrat) research subjects [40].

**Online vs. Offline:** Several researchers have studied the advantages and disadvantages of web-based vs. telephone or other traditional survey methodologies [2, 23, 30, 86, 90, 105, 107], with Dillman [21] providing a book-length overview. Sinclair *et al.* [86] focus on epidemiological research, which frequently requires collecting data from a representative sample of the community, or recruiting members of specific groups through broad community approaches. They look at response rates for mail and telephone surveys, but web surveys they consider involve direct mailing of postcards and inviting recipients to fill out an online survey and as such do not provide compelling incentives compared to crowd-sourced studies. Fricker [30] compare telephone and Web versions of a questionnaire that assessed attitudes toward science and knowledge of basic scientific facts. However, again, the setting differs significantly from that of InterPoll, in that crowd workers have a direct incentive to participate and complete the surveys.

Duffy [23] give a comparison of online and face-to-face surveys. Issues studies include interviewer effect and social desirability bias in face-to-face methodologies; the mode effects of online and face-to-face survey methodologies, including how response scales are used; and differences in the profile of online panelists, both demographic and attitudinal. Interestingly, Duffy *et al.* report questions pertaining to technology use should not be asked online, as they result in much higher use numbers (i.e., *PC use at home* is 91% in the online sample vs. 53 in the face-to-face sample). Surprisingly, these differences pertain even to technologies such as DVD players and digital TV. They also conclude that online participants are frequently better informed about issues such as cholesterol, and are likely to quickly search for an answer, which compromises the ability to ask knowledge-based questions, especially in a crowd setting. Another conclusion is that for online populations, propensity score weighting has a significant effect, especially for politically-oriented questions.

Stephenson *et al.* [90] study the validity of using online surveys vs. telephone polls by examining the differences and similarities between parallel Internet and telephone surveys conducted in Quebec after the provincial election in 2007. Both samples have demographic characteristics differing slightly, even after re-weighting, from that of the overall population. Their results indicate that the responses obtained in each mode differ somewhat, but that few inferential differences would occur depending on which dataset were used, highlighting the attractiveness of online surveys, given their generally lower cost.

**Biases:** Biases in online crowds, compared to online populations, general populations as well as population samples obtained via different recruitment techniques have attracted a lot of attention [3, 47, 73, 74, 76, 75, 80], but most conclusions have been positive. In particular, crowds often provide more diversity of participants, on top of higher completion rates and frequently quality of work.

Antin *et al.* [3] study the *social desirability bias* on MECHANICAL TURK. They use a survey technique called *the list experiment* which helps to mitigate the effect of social desirability on survey self-reports. Social desirability bias refers to "the tendency of people to deny socially undesirable traits or qualities and to admit to socially desirable ones" [15]. Among US Turkers, they conclude that social desirability encourages over-reporting of each of four motivating factors examined; the over-reporting was particularly large in the case of money as a motivator. In contrast, among Turkers in India we find a more complex pattern of social desirability effects, with workers under-reporting "killing time" and "fun" as motivations, and drastically over-reporting "sense of purpose."

**Survey sites:** In the last several years, we have seen surveys sites that are crowd-backed. The key distinction between these sites and INTERPOLL is our focus on optimizations and statistically significant results at the lowest cost. In contrast, survey sites generally are incentivized to encourage the survey-maker to solicit as many participants as possible . At the same time, we draw inspiration from many useful features that the sites described below provide.

Most survey cites give easy access to non-probability samples of the Internet population, generally without attempting to correct for the inherent population bias. Moreover, while Internet use in the United States is approaching 85% of adults, users tend to be younger, more educated, and have higher incomes [72]. Unlike other tools we have found, Google Customer Surveys support re-weighting the survey results to match the demographics of the Current Population Survey (CPS) [100].

SurveyMonkey claims to be the most popular survey building platform [41]. In recent years, they have added support for data analytics as well as an on-demand crowd. Market research seems to be the niche they are trying to target [92]. SurveyMonkey performs ongoing monitoring of audience quality through comparing the answers they get from their audience to that obtained via daily Gullop telephone polls [91]. They conclude that the SurveyMonkey Audience 3-day adjusted average, for 5 consecutive days is within a 5% error margin of Gallup's 14-day trailing average. In other words, when corrected for a higher average income of SurveyMonkey respondents in comparison to the US census data, SurveyMonkey is able to produce effectively the same results as Gallup, with only 3-days of data instead of 14 for Gallup.

Instant.ly and uSamp [101] focus primarily on marketing studies and boast an on-demand crowd with very fast turn-around times: some survey are completed in minutes. In addition to rich demographic data, uSamp collects information on the industry in which respondents are employed, their mobile phone type, job title, etc., also allowing one to filter and aggregate using these demographic characteristics.

Unlike other sites, Google Surveys results have been studied in academic literature. McDonald *et al.* [66] compares the responses of a probability based Internet panel, a non-probability based Internet panel, and Google Consumer Surveys against several media consumption and health benchmarks, leading the authors to conclude that despite differences in survey methodology, Consumer Surveys can be used in place of more traditional Internet-based panels without sacrificing accuracy.

Keeter *et al.* [48] present a comparison of results performed at Pew to those obtained via Google Customer Surveys. Note that demographic characteristics for survey-takes appear to be taken from DoubleClick cookies and are generally inferred and not verified (an approach taken by Instant.ly). A clear advantage of this approach is asking fewer questions; however, there are obvious disadvantages.

Apparently, for about 40% of survey-takes, reliable demographic information cannot be determined. The Google Consumer Survey method samples Internet users by selecting visitors to publisher websites that have agreed to allow Google to administer one or two questions to their users. As of 2012, there are about 80 sites in the *Google Surveys publisher network* (and 33 more currently in testing). The selection of surveys for eligible visitors of these sites appears random. Details on the Google Surveys "survey wall" appear scarce [24].

The Pew study attempted to validate the inferred demographic characteristic and concluded that for 75% of respondents, the inferred gender matched their survey response. For age inference, the results were mixed, with about 44% confirming the automatically inferred age range. Given that the demographic characteristics are used to create a stratified sample, and to re-weight the survey results, these differences may lead to significant errors; for instance, fewer older people using Google Consumer Surveys approved of Obama's job performance than in the Pew Research survey. The approach taken in InterPoll is to *ask* the user to provide their demographic characteristics; we would immensely benefit from additional support on the back-end level to obtain or verify the user-provided data. Google Customer Surveys have been used for information political surveys [55].

The Pew report concludes that, demographically, the Google Consumer Surveys sample appears to conform closely to the demographic composition of the overall internet population. From May to October 2012, the Pew Research Center compared results for 48 questions asked in dual frame telephone surveys to those obtained using Google Consumer Surveys. Questions across a variety of subject areas were tested, including: demographic characteristics, technology use, political attitudes and behavior, domestic and foreign policy and civic engagement. Across these various types of questions, the median difference between results obtained from Pew Research surveys and using Google Consumer Surveys was 3 percentage points. The mean difference was 6 points, which was a result of several sizable differences that ranged from 10–21 points and served to increase the mean difference. It appears, however, that Google Survey takers are no more likely to be technology-savvy than an average Internet user, largely eliminating that bias. A key limitation for large-scale survey appears to be the inability to ask more than a few questions at a time, which is a limitation of their format [24], and the inability to administer questions to the same responder over time. The focus in InterPoll is on supporting as many questions as the developer wants to include.

## 4 Conclusions

This paper presents a vision for InterPoll, a language integrated approach to programming crowd-sourced polls. While much needs to be done to achieve the goals outlined in Section 1, we envision InterPoll as a powerful system, useful in a range of domains, including social sciences, political and marketing polls, and health surveys.

## References

1 Sarah Anderson, Sarah Wandersee, Ariana Arcenas, and Lynn Baumgartner. Craigslist samples of convenience: recruiting hard-to-reach populations. Unpublished.

2 D Andrews, B Nonnecke, and J Preece. Electronic survey methodology: A case study in reaching hard-to-involve Internet users. *International Journal of ...*, 2003.

3 J Antin and A Shaw. Social desirability bias and self-reports of motivation: a study of Amazon Mechanical Turk in the US and India. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

4 Daniel Barowy, Charlie Curtsinger, Emery Berger, and Andrew McGregor. AutoMan: A platform for integrating human-based and digital computation. *Proceedings of the ACM international conference on Object oriented programming systems languages and applications – OOPSLA'12*, page 639, January 2012.

5 T S Behrend, D J Sharek, and A W Meade. The viability of crowdsourcing for survey research. *Behavior research methods*, January 2011.

6 A Berinsky, G Huber, and G Lenz. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368, July 2012.

7 Adam J AJ Berinsky, Gregory A GA Huber, and Gabriel S Lenz. Using mechanical Turk as a subject recruitment tool for experimental research. *Typescript, Yale*, pages 1–26, 2010.

8 Samuel J. Best and Brian S. Krueger. *Exit Polls: Surveying the American Electorate, 1972–2010.* CQ Press, 2012.

9 M Bleja, T Kowalski, and K Subieta. Optimization of object-oriented queries through rewriting compound weakly dependent subqueries. *Database and Expert Systems*, pages 1–8, January 2010.

10 James Bornholt, Todd Mytkowicz, and Kathryn S. McKinley. Uncertain<T>: A First-order Type for Uncertain Data. *SIGARCH Comput. Archit. News*, 42(1):51–66, 2014.

11 M Buhrmester and T Kwang. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *on Psychological Science*, January 2011.

12 Trent D Buskirk, D Ph, and Charles Andrus. Online Surveys Aren't Just for Computers Anymore! Exploring Potential Mode Effects between Smartphone and Computer-Based Online Surveys. *American Statistical Association (ASA), events and resources for statisticians, educators, students*, pages 5678–5691, 2010.

13 J Chandler, P Mueller, and G Paolacci. Methodological concerns and advanced uses of crowdsourcing in psychological research. *Behavioral Research*, 2013.

14 James Cheney, Sam Lindley, and Philip Wadler. A practical theory of language-integrated query. *Proceedings of the 18th ACM SIGPLAN international conference on Functional programming – ICFP'13*, page 403, January 2013.

15 D. L. Clancy and K. Phillips J. Some effects of "Social desirability" in survey studies. *The American Journal of Sociology*, 77(5):921–940, 1972.

16 Christopher Cooper, David M McCord, and Alan Socha. Evaluating the college sophomore problem: the case of personality and politics. *Journal of Psychology*, 145(1):23–37, 2011.

17 M Couper. Designing effective web surveys, 2008.

18 M P Couper. Review: Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, pages 1–31, January 2000.

19 Franco Curmi and Maria Angela Ferrario. Online sharing of live biometric data for crowd-support: Ethical issues from system design. Unpublished, 2013.

20 Nilesh Dalvi, Christopher Ré, and Dan Suciu. Probabilistic Databases: Diamonds in the Dirt. *Communications of the ACM*, 2009.

21 D Dillman, R Tortora, and D Bowker. Principles for constructing Web surveys. Unpublished, 1998.

22 M Duda and J Nobile. The fallacy of online surveys: No data are better than bad data. *Human Dimensions of Wildlife*, 2010.

23 B Duffy, K Smith, and G Terhanian. Comparing data from online and face-to-face surveys. *International Journal of*, January 2005.

24 Justin Ellis. How Google is quietly experimenting in new ways for readers to access publishers' content, 2011.

25 Joel Evans, New Hempstead, and Anil Mathur. The value of online surveys. *Internet Research*, 15(2):195–219, January 2005.

26 Jeremy Eysenbach, Gunther Eysenbach, and Jeremy Wyatt. Using the Internet for Surveys and Health Research. *Journal of Medical Internet Research*, 4(2):e13, January 2002.

27 Emma Ferneyhough. Crowdsourcing Anxiety and Attention Research, 2012.

28 K Fort, G Adda, and K B Cohen. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, pages 1–8, January 2011.

29 Michael Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. CrowdDB: answering queries with crowdsourcing. *SIGMOD'11: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1–12, June 2011.

30 S Fricker, M Galesic, R Tourangeau, and T Yan. An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 2005.

31 M Fuchs. Mobile Web Survey: A preliminary discussion of methodological implications. *Envisioning the survey interview of the future*, January 2008.

32 Marek Fuchs and Britta Busse. The Coverage Bias of Mobile Web Surveys Across European Countries. *International Journal of Internet Science*, 4(1):21–33, 2009.

33 Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.

34 Andrew D Gordon, Johannes Borgstr, Nicolas Rolland, and John Guiver. Tabular: A Schema-Driven Probabilistic Programming Language. Technical report, Microsoft Research, 2013.

35 Samuel Gosling, Simine Vazire, Sanjay Srivastava, and Oliver John. Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59(2):93–104, January 2004.

36 R.M. Groves. *Survey Errors and Survey Costs*. Wiley Series in Probability and Statistics. Wiley, 1989.

37 Robert M. Groves, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology*. Wiley, 2009.

38 Torsten Grust, Jan Rittinger, and Tom Schreiber. Avalanche-safe LINQ compilation. *Proceedings of the VLDB Endowment*, 3(1-2):162–172, September 2010.

39 H Gunn. Web-based surveys: Changing the survey process. *First Monday*, 2002.

40 Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *The Behavioral and brain sciences*, 33(2-3):61–83; discussion 83–135, June 2010.

41 HubSpot and SurveyMonkey. Using online surveys in your marketing. Unpublished.

42 P G Ipeirotis. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads*, January 2010.

43 P G Ipeirotis. Demographics of Mechanical Turk. *2010*, January 2010.

44 Floyd J. Fowler Jr. *Survey Research Methods (4th ed.)*. SAGE Publications, Inc., 4 edition, 2009.

45 R Jurca and B Faltings. Incentives for expressing opinions in online polls. *Proceedings of the ACM Conference on Electronic Commerce*, 2008.

46 Adam Kapelner and Dana Chandler. Preventing Satisficing in Online Surveys : A "Kapcha" to Ensure Higher Quality Data. *CrowdConf*, 2010.

47 S Keeter. The impact of cell phone noncoverage bias on polling in the 2004 presidential election. *Public Opinion Quarterly*, 2006.

48 Scott Keeter, Leah Christian, and Senior Researcher. A Comparison of Results from Surveys by the Pew Research Center and Google Consumer Surveys. http://www.people-press.org/files/legacy-pdf/11-7-12 Google Methodology paper.pdf, 2012.

49 P Kellner. Can online polls produce accurate findings? *International Journal of Market Research*, 2004.

50 A Kittur, E H Chi, and B Suh. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the SIGCHI conference on*, January 2008.

51 Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. CrowdWeaver: Visually Managing Complex Crowd Work. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work – CSCW'12*, page 1033, January 2012.

52 R Kosara and C Ziemkiewicz. Do Mechanical Turks dream of square pie charts? *Proceedings Beyond time and errors: novel evaLuation methods for Information Visualization*, 2010.

53 Robert Kraut, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen, and Mick Couper. Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, 59(2):105–117, January 2004.

54 Robert E Kraut. CrowdForge : Crowdsourcing Complex Work. *UIST*, pages 43–52, 2011.

55 Paul Krugman. What People (Don't) Know About The Deficit, April 2013.

56 A Kulkarni, M Can, and B Hartmann. Collaboratively crowdsourcing workflows with turkomatic. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, January 2012.

57 A P Kulkarni, M Can, and B Hartmann. Turkomatic: automatic recursive task and workflow design for mechanical turk. *CHI'11 Extended Abstracts on Human*, January 2011.

58 G Little, L B Chilton, M Goldman, and R C Miller. TurKit: tools for iterative tasks on Mechanical Turk. *Proceedings of UIST*, pages 1–2, January 2009.

59 Benjamin Livshits and George Kastrinis. Optimizing human computation to save time and money. Technical Report MSR-TR-2014-145, Microsoft Research, November 2014.

60 Benjamin Livshits and Todd Mytkowicz. Saving money while polling with interpoll using power analysis. In *In Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP 2014)*, November 2014.

61 A Marcus, E Wu, Karger, S R Madden, and R C Miller. Crowdsourced databases: Query processing with people. *2011*, January 2011.

62 Adam Marcus, David Karger, Samuel Madden, Robert Miller, and Sewoong Oh. Counting with the crowd. *Proceedings of the VLDB Endowment* ,, 6(2), December 2012.

63 Adam Marcus, Eugene Wu, David Karger, Samuel Madden, and Robert Miller. Human-powered sorts and joins. *Proceedings of the VLDB Endowment* ,, 5(1), September 2011.

64 W Mason and S Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, January 2012.

65 Joe Mayo. *LINQ Programming*. McGraw-Hill Osborne Media, 1 edition, 2008.

66 Paul Mcdonald, Matt Mohebbi, and Brett Slatkin. Comparing Google Consumer Surveys to Existing Probability and Non-Probability Based Internet Surveys. `http://www.google.com/insights/consumersurveys/static/consumer_surveys_whitepaper.pdf`.

67 Patrick Minder, Sven Seuken, Abraham Bernstein, and Mengia Zollinger. CrowdManager – Combinatorial Allocation and Pricing of Crowdsourcing Tasks with Time Constraints. *Workshop on Social Computing and User Generated Content in conjunction with ACM Conference on Electronic Commerce (ACM-EC 2012)*, 2012.

68 Derek Murray, Michael Isard, and Yuan Yu. Steno: automatic optimization of declarative queries. *Proceedings of the Conference on Programming Language Design and Implementation*, pages 1–11, June 2011.

69 Venkata Nerella, Sanjay Madria, and Thomas Weigert. An Approach for Optimization of Object Queries on Collections Using Annotations. *2013 17th European Conference on Software Maintenance and Reengineering*, pages 273–282, March 2013.

70 Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, July 2009.

71 G Paolacci, J Chandler, and P G Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision*, January 2010.

72 Pew Research Center. Demographics of Internet users, 2013.

73 Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications : a publication of the Ecological Society of America*, 19(1):181–97, January 2009.

74 P Podsakoff, S MacKenzie, and J Lee. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879–903, 2003.

75 Ramo and S M Hall. Reaching young adult smokers through the Internet: Comparison of three recruitment mechanisms. *Nicotine & Tobacco*, January 2010.

76 D Ramo, S Hall, and J Prochaska. Reliability and validity of self-reported smoking in an anonymous online survey with young adults. *Health Psychology*, 2011.

77 Allan Roshwalb, Neal El-Dash, and Clifford Young. Toward the use of Bayesian credibility intervals in online survey results. `http://www.ipsos-na.com/knowledge-ideas/public-affairs/points-of-view/?q=bayesian-credibility-interval`, 2012.

78 J Ross, A Zaldivar, L Irani, B Tomlinson, and M Silberman. Who are the crowdworkers?: shifting demographics in Mechanical Turk. *CHI'10 Extended*, January 2009.

79 Matthew Salganik and Karen Levy. Wiki surveys: Open and quantifiable social data collection. http://arxiv.org/abs/1202.0500, February 2012.

80 L Sax, S Gilmartin, and A Bryant. Assessing response rates and nonresponse bias in web and paper surveys. *Research in higher education*, 2003.

81 L Schmidt. Crowdsourcing for human subjects research. *Proceedings of CrowdConf*, 2010.

82 M Schonlau, A Soest, A Kapteyn, and M Couper. Selection bias in Web surveys and the use of propensity scores. *Sociological Methods & Research*, 37(3):291–318, February 2009.

83 G Schueller and A Behrend. Stream Fusion using Reactive Programming, LINQ and Magic Updates. *Proceedings of the International Conference on Information Fusion*, pages 1–8, January 2013.

84 S Sills and C Song. Innovations in survey research an application of web-based surveys. *Social science computer review*, 2002.

85 Cindy D. Simasa2 and Elizabeth N. Kama. Risk Orientations and Policy Frames. *The Journal of Politics*, 72(2), 2010.

86 Martha Sinclair, Joanne O'Toole, Manori Malawaraarachchi, and Karin Leder. Comparison of response rates and cost-effectiveness for a community-based survey: postal, internet and telephone modes with generic or personalised recruitment approaches. *BMC medical research methodology*, 12(1):132, January 2012.

87 Nick Sparrow. Developing Reliable Online Polls. *International Journal of Market Research*, 48(6), 2006.

88 Robin Sprou. Exit Polls: Better or Worse Since the 2000 Election? Joan Shorestein Center on the Press, Politics and Public Policy, 2008.

89 J Sprouse. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, January 2011.

90 L B Stephenson and J Crête. Studying political behavior: A comparison of Internet and telephone surveys. *International Journal of Public Opinion Research*, January 2011.

91 SurveyMonkey. Data Quality: Measuring the Quality of Online Data Sources. http://www.slideshare.net/SurveyMonkeyAudience/surveymonkey-audience-data-quality-whitepaper-september-2012, 2012.

92 SurveyMonkey. Market Research Survey; Get to know your customer, grow your business, 2013.

93 M Swan. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research*, January 2012.

94 Melanie Swan. Scaling crowdsourced health studies : the emergence of a new form of contract research organization. *Personalized Medicine*, 9:223–234, 2012.

95 Swati Tawalare and S Dhande. Query Optimization to Improve Performance of the Code Execution. *Computer Engineering and Intelligent Systems*, 3(1):44–52, January 2012.

96 Emma Tosch and Emery D. Berger. Surveyman: Programming and automatically debugging surveys. In *Proceedings of Conference on Object Oriented Programming Systems Languages and Applications*, OOPSLA'14, 2014.

97 Roger Tourangeau, Frederick G. Conrad, and Mick P. Couper. *The Science of Web Surveys*. Oxford University Press, 2013.

98 H L Truong, S Dustdar, and K Bhattacharya. Programming hybrid services in the cloud. *Service-Oriented Computing*, pages 1–15, January 2012.

99 Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458, 1981.

100 US Census. Current population survey, October 2010, school enrollment and Internet use supplement file, 2010.

101 USamp. Panel Book 2013. 2013.

102 A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 06 1945.

103 Fabian L Wauthier and Michael I Jordan. Bayesian Bias Mitigation for Crowdsourcing. *Neural Information Processing Systems Conference*, pages 1–9, 2011.

104 R W White. Beliefs and Biases in Web Search. *2013*, January 2013.

105 K Wright. Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, 2005.

106 J Wyatt. When to use web-based surveys. *Journal of the American Medical Informatics Association*, 2000.

107 D Yeager, J Krosnick, L Chang, and H Javitz. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 2011.

108 X Yin, W Liu, Y Wang, C Yang, and L Lu. What? How? Where? A Survey of Crowdsourcing. *Frontier and Future Development of*, January 2014.

109 Clifford Young, John Vidmar, Julia Clark, and Neale El-Dash. Our brave new world: blended online samples and performance of no probability approaches. Ipsos Public Affairs.