Dynamic Question Ordering: Obtaining Useful Information While Reducing User Burden

Kirstin Early

August 2017
CMU-ML-17-103

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jennifer Mankoff, Co-chair
Stephen Fienberg, Co-chair
Jared Murray
Barnabás Póczos
John Abowd, Cornell University & U.S. Census Bureau

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Machine Learning

This research was sponsored by the National Science Foundation under grant numbers IIS1217929, IIS1320402, and SES1130706, the Bureau of Census grant number YA132317SE0096, and the University of California Berkeley Foundation Siebel Seed funding.

The statistical results in this document were produced as part of an internal U.S. Census Bureau research project DMS 1262 supervised by the American Community Survey Office. The results have been examined to insure that they include no confidential data and were released as part of clearance requests DMS1262-20170721 and DMS1262-20170803. Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Keywords: Adaptive data collection, cost-effective data collection, machine learning, adaptive survey design, survey methodology, questionnaire design.

Abstract

As data become more pervasive and computing power increases, the opportunity for transformative use of data grows. Collecting data from individuals can be useful to the individuals (by providing them with personalized predictions) and the data collectors (by providing them with information about populations). However, collecting these data is costly: answering survey items, collecting sensed data, and computing values of interest deplete finite resources of time, battery life, money, *etc.* Dynamically ordering the items to be collected, based on already known information (such as previously collected items or paradata), can lower the costs of data collection by tailoring the information-acquisition process to the individual. This thesis presents a framework for an iterative dynamic item ordering process that trades off item utility with item cost at data collection time. The exact metrics for utility and cost are application-dependent, and this framework can apply to many domains. The two main scenarios we consider are (1) data collection for personalized predictions and (2) data collection in surveys. We illustrate applications of this framework to multiple problems ranging from personalized prediction to questionnaire scoring to government survey collection. We compare data quality and acquisition costs of our method to fixed order approaches and show that our adaptive process obtains results of similar quality at lower cost.

For the personalized prediction setting, the goal of data collection is to make a prediction based on information provided by a respondent. Since it is possible to give a reasonable prediction with only a subset of items, we are not concerned with collecting all items. Instead, we want to order the items so that the user provides information that most increases the prediction quality, while not being too costly to provide. One metric for quality is prediction certainty, which reflects how likely the true value is to coincide with the estimated value. Depending whether the prediction problem is continuous or discrete, we use prediction interval width or predicted class probability to measure the certainty of a prediction. We illustrate the results of our dynamic item ordering framework on tasks of predicting energy costs, student stress levels, and device identification in photographs and show that our adaptive process achieves equivalent error rates as a fixed order baseline with cost savings up to 45%.

For the survey setting, the goal of data collection is often to gather information from a population, and it is desired to have complete responses from all samples. In this case, we want to maximize survey completion (and the quality of necessary imputations), and so we focus on ordering items to engage the respondent and collect hopefully all the information we seek, or at least the information that most characterizes the respondent so imputed values will be accurate. One item utility metric for this problem is information gain to get a "representative" set of answers from the respondent. Furthermore, paradata collected during the survey process can inform models of user engagement that can influence either the utility metric (*e.g.*, likelihood the respondent will continue answering questions) or the cost metric (*e.g.*, likelihood the respondent will break off from the survey). We illustrate the benefit of dynamic item ordering for surveys on two nationwide surveys conducted by the U.S. Census Bureau: the American Community Survey and the Survey of Income and Program Participation.

To Spencer, my brother and favorite person.

Acknowledgments

So many people have helped me during my time at CMU and my work on this thesis. I have been very fortunate to know you all!

First, the work in this thesis and my development as a researcher have been immensely influenced by my incredible advisors, Jen Mankoff and Steve Fienberg. The benefits of collaborating across three departments and busy schedules far outweighed the logistical challenges (like getting us all in the same room). I aspire to Jen's research approach of centering real-world problems and devising solutions that are both innovative and meaningful, and Steve's ability to make connections across seemingly disparate ideas and fields. Jen is committed to making sure her students are doing well in life, as well as research, and Steve's extreme confidence in his students has been motivating and reassuring. I appreciate the mentorship and encouragement from both of them!

The rest of my thesis committee—John Abowd, Jared Murray, and Barnabás Póczos—also provided helpful guidance on my research. Though not part of my committee, I received useful feedback on my work from Bill Eddy, Rebecca Nugent, and Sam Ventura.

In addition to John, a number of people at the Census Bureau have been key in making sure my thesis work happened. Nancy Bates and Renee Ellis coordinated the NSF-Census Research Network and helped me get access to Census data for a non-traditional research project. Sandy Clarke and Dave Raglin answered my many questions about the format and meaning of the American Community Survey questions and paradata. Holly Monti and Lori Reeder led an informative workshop on the SIPP Synthetic Beta in August 2016 that helped me make use of that dataset.

I've gotten to work with lots of other people on projects that are not part of this thesis and have learned a lot from their skill sets and approaches to research. Studying the interplay of gender and authorship has been a blast with Jen, Jess Hammer, Jen Rode, Veronica Cucuiat, Houda El Mimouni, Meg Richards, and Anna Wong. I also learned a lot working on EDigs with Jen, Dimeji Onafuwa, Vikram Kamath, Aleks Tapinsh, and Nidhi Vyas.

Nothing in any department would work if not for the administrative staff. I am very grateful for Diane Stidle, Marian D'Amico, Carloz Gil, and Margie Smykla, who perform a multitude of tasks for us students: arranging reimbursements, filling out forms, registering us for classes, and generally making sure everything is running smoothly.

I have enjoyed being part of several groups that promote diversity at SCS and CMU. Thanks so much to Carol Frieze, Mary Widom, Ashley Grice, and Carrie Hagan for their work in building these communities and making CMU a more welcoming place!

My friends have made my time in Pittsburgh not only intellectually rewarding but also a lot of fun. Notable MLD friends include Nicole Rafidi, Calvin Murdock, Mariya Toneva, Junier Oliva (the funniest person in the department), Avinava Dubey, Micol Marchetti-Bowick, Willie Neiswanger, Maria de Arteaga, Anthony Platanios, Mrinmaya Sachan, Dan Howarth, and Dan Schwartz. Thanks for making MLD the best department! I've also had an excellent time with the swim crew (Nicole, Erin McCormick, Rony Patel, and Milda Zizyte), gym posse (Milda, Erin, Rony, and Thom Popovici), and Jonathan Porras and our many mutual friends. Kayla Frisoli and I spent many hours together in cars, hotel rooms, and the Census Bureau this past semester, which did end up being fun.

Finally, thanks to my family for always encouraging my education and never asking why I've been in school for this long. Living in Pennsylvania offered opportunities to see some formerly faraway family more frequently, like my aunt and uncle, Cindy and Bill Pierce (also known as my Thanksgiving hosts for the past five years). It's been especially fun to have Spencer do his undergrad at CMU while I've been here!

I've definitely accidentally omitted some people—as long as you're not the guy who semi-regularly plays bagpipes underneath my office window at 8 am, you have probably enriched my CMU experience, so thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online data collection from individuals can provide them with personalized predictions (*e.g.*, recommender systems), in addition to gathering information from populations of interest (*e.g.*, surveys), at scale and at low cost to the data collectors. However, users do not have the resources to provide all the information we seek—they cannot answer too many questions that may be difficult to get answers for, and their mobile devices do not have sufficient battery life to provide constant streams of high-fidelity sensed data. Strategically choosing which feature to obtain next from a particular user, depending on previous responses, can lower the burden on respondents while still collecting useful information. This idea of dynamic question ordering (DQO) can also take into account the varying *costs* of individual questions—a question can be costly to answer due to the effort required for a respondent to come up with an answer, the drain on a battery for sensing a certain measurement, or the likelihood that a respondent will break off from a survey when presented with the question (among other types of cost). DQO trades off the utility we get from having an answer to a question against its cost and sequentially requests items in order to make useful, confident predictions and gather survey data with the resources users are willing and able to provide.

We propose a general framework for dynamically ordering questions, based on previously collected data, to engage respondents, improving prediction quality and survey completion. Our work considers two scenarios for data collection from survey-takers. In the first, we want to give the respondent a personalized prediction, based on information they provide. Since it is possible to give a reasonable prediction with only a subset of questions, we are not concerned with motivating the user to answer all questions. Instead, we want to order questions so that the user provides information that most reduces the uncertainty of our prediction, while not being too burdensome to answer. In the second scenario, our goal is to maximize survey completion (and the quality of necessary imputations) and so we focus on ordering questions to engage the respondent and collect hopefully all the information we seek, or at least the information that most characterizes the respondent so imputed values will be accurate.

1.1 A note on vocabulary

This idea of dynamic question ordering can be applied to a broad number of applications and domains, and the vocabulary is often specific to each domain, with different terms capturing similar concepts across applications. In this thesis, rather than enforcing a standard vocabulary across all domains, we will adopt the domain-specific terminology for each application, keeping in mind that the specific word refers to a more general idea. Table 1.1 summarizes these equivalent terms, according to which main domain (prediction or survey-taking) they appear in.

Table 1.1: Equivalent words used for different applications in this thesis.

Meaning	Prediction-focused	Survey-focused
Information to be collected	Feature	Question
Time of data collection	Test time	Survey time
Information provider	User/system	Respondent
Information cost	Cost	Burden
Method name	FOCUS	DQO

1.2 Thesis statement and contributions

Dynamically ordering questions, based on already known information (such as previously answered questions or paradata), can lower the costs of data collection and prediction by tailoring the information-acquisition process to the individual. This thesis presents a framework for an iterative dynamic question ordering process that trades off question utility against question cost at data collection time. The exact metrics for utility and cost are application-dependent. We compare data quality and acquisition costs of our method to fixed order approaches and show that our adaptive process obtains results of similar or better quality at lower cost.

1.2.1 Contributions

The contributions of this thesis center around the development and evaluation of a framework for personalized, adaptive data collection for obtaining useful information at low user burden. In this framework, information is gathered sequentially. At each step of data collection, the information collected up to that point is used in determining which piece of information will be most helpful, for the particular problem at hand. Thus, the order in which items are acquired is determined at data collection time and is likely to differ for each instance of data collection.

First, we develop a general framework for cost-effective dynamic question ordering (DQO) that trades off question utility against question cost in sequential data collection. Both question utility and question cost are determined in the context of what information has been collected so far. Therefore, different pieces of information can provide different levels of usefulness to the purpose of the data collection at different cost values at different points during data collection. This broad framework encompasses much prior work in adaptive data collection methods in the fields of machine learning, educational testing and psychometrics, and survey methodology.

Next, we develop a variety of problem-specific definitions of question utility and question cost to apply the DQO framework to concrete problems. Because the DQO framework is abstractly defined in terms of utility and cost, it is general enough to apply to many scenarios. Therefore, the utility and cost metrics need to be defined before DQO can be applied in practice. Examples of question utility metrics include prediction quality, information, and respondent engagement. Examples of question cost metrics include user burden, item nonresponse rate, and item response time.

Finally, we apply the specific instantiations of the DQO framework to several domains: test-time feature acquisition, with applications in residential energy analysis, health, and ubiquitous computing; questionnaire scoring, with applications in measuring cultural values; and survey data collection, with applications in two nationwide surveys conducted by the U.S. Census Bureau. Our experiments illustrate that DQO achieves similar- or better-quality outcomes (measured in terms of prediction accuracy and certainty, imputation quality, and survey completion rates) than fixed-order approaches to data collection, while being less costly to respondents.

1.3 Outline

The remainder of this dissertation is organized as follows: Chapter 2 covers previous work on adaptive data collection, in the areas of machine learning, adaptive testing, and survey methodology. Chapter 3 identifies what gaps remain in the past work and presents a general framework for dynamic question ordering. Chapters 4, 5, and 6 tailor this framework to situations of prediction, questionnaire scoring, and survey-taking, respectively, along with example applications in domains such as energy, health, ubiquitous computing, political science, and government surveys. Finally, Chapter 7 concludes the dissertation and gives directions for future work.

Chapter 2

Related work

There is a rich literature focusing on adaptively ordering acquisition of information to improve outcomes while minimizing costs, across multiple fields. For example, test-time feature acquisition in machine learning considers the problem of adaptively gathering the most influential feature values to make a prediction on a new test point, whose features can only be acquired at cost. Adaptive testing in educational research and psychometrics looks at a similar problem of making accurate measurements of an examinee's skill level using few questions. Adaptive survey design in survey methodology considers the broader problem of gathering good-quality data from a population that accurately represents that population.

The following sections summarize past work in these areas and discuss how the work in this thesis advances prior work.

2.1 Test-time feature acquisition

In standard supervised machine learning approaches, a predictor is learned from training data and used to make a prediction on a test point. It is assumed that the training data and test points share a common set of features, but labels are provided for only the training data.

Because labels can be costly to acquire at training time, a large body of related work has focused on active learning, which strategically selects which unlabeled data to acquire labels for, so as to maximize a model's performance while minimizing the cost of data collection (Cohn, Ghahramani, & Jordan, 1996).

Alternatively, it may be features that are costly to acquire at test time (or, equivalently, prediction time, for deployed systems). For application areas where feature computation time is a bottleneck at test time (*e.g.*, natural language processing (NLP), computer vision), the primary goal of test-time feature acquisition is to speed up prediction (*e.g.*, (He, Daumé III, & Eisner, 2013; Strubell, Vilnis, Silverstein, & McCallum, 2015; Weiss & Taskar, 2013)). In other domains, there is an allowable test-time budget of costs other than time, such as user burden of providing information (*e.g.*, (Early, Mankoff, & Fienberg, 2017; Early, Fienberg, & Mankoff, 2016)) or the resulting loss in privacy from disclosing information (*e.g.*, (Pattuk, Kantarcioglu, Ulusoy, & Malin, 2015)).

Feature selection—*i.e.*, choosing at training time a relevant subset of features to include in a model (Guyon & Elisseeff, 2003)—can cut down on the number of features that must be acquired to make a prediction on a new instance at test time. However, typically, the main motivations for feature selection are (1) avoiding overfitting to the training set in high-dimensional problems and (2) generating interpretable models. Some past work (*e.g.*, (Ballesteros & Bohnet, 2014)) has approached training-time feature selection with the goal of reducing test-time prediction costs. Other researchers have pointed out that the optimal budget-constrained set and/or number of features to acquire likely depends on each particular instance. Thus, there is a need for test-time, dynamic feature selection.

2.1.1 Cost and order at training time; number at test time

One approach to test-time feature selection is learning sequences of features to add, from training data, and then acquiring features in this order at test time. For example, Strubell et al. (2015) learn an ordering

of features at training time through the use of prefix scores that capture the prediction margin (*i.e.*, how much more confident the classifier is in the correct prediction than any other). At test time, they then acquire features in this order, computing prefix scores at each stage until some label is predicted above all others by a specified margin. They apply this technique to several problems in NLP (part-of-speech tagging, dependency parsing, and named-entity recognition). A related method is classifier cascades and trees (Xu, Kusner, Weinberger, Chen, & Chapelle, 2014), which learn at training time cost-sensitive trees or cascades, where each node is a classifier. At test time each instance is passed through the tree or cascade, evaluating additional features as necessitated by the tree structure.

While these approaches dynamically decide how many features to acquire in an instance-specific fashion at test time, they do not determine an instance-specific order in which features are acquired at test time. As a result, such methods may have to obtain more features than would be necessary to adequately predict any given instance, just to get the relevant features for that particular instance.

2.1.2 Cost at training time; number and order at test time

A more flexible approach is to decide both order and number of features to acquire at test time. Across several domains, test-time feature selection has been modeled as a Markov decision process (MDP) (*e.g.*, (He, Daumé III, & Eisner, 2012; He et al., 2013; Samadi, Talukdar, Veloso, & Mitchell, 2015; Shi, Steinhardt, & Liang, 2015; Weiss & Taskar, 2013)). Generally, these methods consider the set of features acquired thus far to be the states of the MDP, the decision of which feature to acquire next as the action, with a reward function that reflects how much the inclusion of the next feature improves the prediction (potentially with a penalty on feature cost). They then learn a policy that chooses which action to take (*i.e.*, feature to add) based on the current state (*i.e.*, current known set of features), from a set of training data (*e.g.*, (He et al., 2012, 2013; Samadi et al., 2015; Weiss & Taskar, 2013)).

This is a fairly general approach that has been used quite widely. For example, Shi et al. (2015) take a similar approach to the problem of constructing heterogeneous sampling algorithms, by considering reward as the improvement in conditional log-likelihood of the label given the sample. Similarly, in the domain of recommender systems, the so-called "cold start problem" (Lika, Kolomvatsos, & Hadjiefthymiades, 2014) is improved by eliciting the most relevant preferences from the user to minimize burden (*e.g.*, (Golbandi, Koren, & Lempel, 2011; Sun et al., 2013)). This is often done by learning a decision tree from training data that can be used at prediction time to decide what sequence of ratings to request (*e.g.*, (Golbandi et al., 2011)). A limitation of such approaches is that they use training data to determine how to ask questions (even if the sequence of features depends on previously provided answers). This approach can be inappropriate for a test sample with behavior very different from the training set.

2.1.3 Cost, number, and order at test time

Finally, there are methods that determine a feature order at test time, using the expected quality of the subsequent prediction to decide which feature to acquire next. For example, Pattuk et al. (2015) formulate a privacy-aware dynamic feature selection algorithm for classification that sequentially chooses features for a test instance, according to which will most increase the expected confidence of the next prediction, as long as including that feature does not violate a privacy constraint. This work is most responsive to the test-time situation. However, it does not address regression, and because its cost metric is defined mathematically by the currently known features (*i.e.*, conditional entropy), the method cannot take context-dependent costs into account.

2.1.4 Summary of key attributes

Table 2.1 summarizes these related works and highlights desired qualities in a test-time feature acquisition algorithm. A full consideration of the issues would accommodate a variety of prediction algorithms.

Table 2.1's Column 1 (**Prediction Problem**) shows that algorithms may predict discrete values (classification) or continuous values (regression)—past work has focused on classification alone. In the **Cost Metric** column, we see that algorithms may assume that all features have equal cost, or they may allow different features to have different costs, which we refer to as feature-specific costs. Next, algorithms vary

Table 2.1: A summary of previous work in test-time feature acquisition and our method (FOCUS, top row). The aspects of test-time feature selection we compare are: whether the type of prediction problem is classification (**C**) or regression (**R**); whether the cost metric for features is Feature-Specific (*i.e.*, different features can have different costs; X means no cost metric); what is determined at test time (**C**ost, **O**rder, # of features); the utility function used (**A**ccuracy or **U**ncertainty or **M**argin from next closest class); in what domains the algorithm has been applied.

Paper	**Prediction Problem**	**Cost Metric**	**Test-time**	**Utility Function**	**Domain**
Early et al. (2016)	C,R	FS	C,O,#	U	Ubicomp/ mobile/energy
He et al. (2012)	C	FS	X	M	—
He et al. (2013)	C	X	X	A	NLP
Strubell et al. (2015)	C	X	#	M	NLP
Shi et al. (2015)	C	X	X	A	Structured prediction
Weiss and Taskar (2013)	C	FS	#	A	Structured prediction
Xu et al. (2014)	C	FS	X	A	LTR, MNIST
Samadi et al. (2015)	C	FS	O,#	M	Knowledge-on-demand
Pattuk et al. (2015)	C	FS	O,#	U	Information disclosure
Trapeznikov and Saligrama (2013)	C	FS	#	A	—
Golbandi et al. (2011) Sun et al. (2013)	C	X	X	A	Recommender systems

on what is done at **Test Time** rather than training time. The optimal number of features needed, the order of features, and the cost of features may all be determined at test time. If cost is determined at test time, algorithms may be able to consider context-dependent costs. For example, in a system that makes medical diagnoses and can request tests (*e.g.*, (Ferrucci, Levas, Bagchi, Gondek, & Mueller, 2013)), it will be less costly to request an invasive biopsy if a surgery is already scheduled in that area. None of the related works we identified support context-dependent cost metrics. The **Utility Function** used to determine the value of a feature also varies. Examples for feature "quality" include subsequent prediction accuracy (*e.g.*, (He et al., 2013; Shi et al., 2015; Weiss & Taskar, 2013; Xu et al., 2014)) and subsequent prediction uncertainty (*e.g.*, (Pattuk et al., 2015)). Finally, the **Domain** column shows where each relevant approach was applied.

An ideal solution to test-time feature acquisition would accommodate a variety of prediction algorithms, allow for feature-specific and context-dependent costs, and support multiple options for prediction utility when requesting information to refine a prediction. Delaying the order determination until test time (rather than learning it from training data) is a requirement for context-dependent costs (since they are not known until test time).

2.2 Questionnaire scoring

A popular use of questionnaires is to summarize an individual's responses into a single score, for a variety of applications, such as psychology, education, and health. Often, this single score measures an underlying trait, such as extraversion, intelligence, or cultural values. The questions used to measure the underlying traits are typically repetitive, to ensure coverage of the trait. This fact can make the questionnaire to measure the latent beliefs unnecessarily long, which discourages respondents from answering all questions thoroughly. "Short form" questionnaires choose a subset of questions to ask respondents. This approach uses the same reduced set of questions for each individual, but different questions may be more informative for certain

individuals than others.

The background for this problem setting involves the concept of determining scores from questionnaire responses, ways to reduce the burden of such questionnaires on respondents, ways to personalize the questionnaires to individual respondents, and finally ways to use personalization for cost reduction of questionnaires. This section first briefly reviews what questionnaire scores are, how they are developed, and what they can measure. Questionnaire scores can summarize a respondent's attitudes or measure a hidden trait determined by the respondent's answers to a set of questions. However, it can be burdensome for a respondent to answer potentially many questions.

We discuss past work to reduce the respondent burden of questionnaires through the use of short form questionnaires that aim to measure the same trait as the full form, but with fewer questions. In addition to being burdensome, standard questionnaires that administer all items to all respondents in the same order do not take advantage of the variation among individuals' answers to questions. We describe how adaptive testing for personalized questionnaires adapts a survey instrument to the current respondent, by taking their previous answers (typically summarized as a current estimate) into account when choosing which test question to administer next. Finally, we discuss related work that has combined cost reduction with personalization by considering question-specific costs when adapting data collection to individuals.

2.2.1 About questionnaire scoring

Especially in psychology, it is common to analyze questionnaire responses for an individual by calculating a score or set of scores to measure the respondent's level of an underlying trait or attitude. One popular example of a score of such underlying traits is the set of Big Five personality traits, which were developed and confirmed by a number of research teams (see (Digman, 1990) for a review). Commonly used "long forms" designed to measure levels of these traits include the 240-item NEO Personality Inventory Revised (NEO-PI-R) (Costa & McCrae, 1985), the 100-item Trait Descriptive Adjectives (TDA) (Goldberg, 1992), the 100-item International Personality Item Pool (IPIP) (Goldberg, 1999), and the 44-item Big Five Inventory (BFI) (John, Donahue, & Kentle, 1991).

It is often the case that a single questionnaire will assign multiple scores to the respondent. For example, the Big Five questionnaires measure five personality traits, and a respondent receives five scores—one for each trait. We will use the term *scale* to refer to a group of questions that measure the same trait. The term *score* will refer to the numeric measurement of the trait that the responses to the scale questions yield.

2.2.2 Reducing the burden of questionnaires: Short forms

Due to the limited resources (including time, interest, and compensation) of participants, researchers have sought ways to reduce respondent burden while still measuring a trait of interest. One common approach is to develop a short form questionnaire to administer to participants. For the Big Five personality traits, the creators of the 240-item NEO-PI-R and the 100-item IPIP produced shorter versions of their questionnaires by selecting subsets of items, resulting in the 60-item NEO Five-Factor Inventory (NEO-FFI) (Costa & McCrae, 1989) and the 50-item IPIP Five-Factor Model (IPIP-FFM) (Goldberg, 1999). Other researchers have also reduced the number of items in these scales, either by doing a factor analysis of a full questionnaire (*e.g.*, to select 40 items from the 100-item TDA (Saucier, 1994), or to select 20 items from the 50-item IPIP-FFM (Donnellan, Oswald, Baird, & Lucas, 2006)) or by hand-selecting a subset of items from each scale (*e.g.*, choosing one positive- and one negative-weighted item from each scale, from the items in the BFI (Rammstedt & John, 2007) or from a combination of the BFI and the TDA (Gosling, Rentfrow, & Swann, 2003)). In the most extreme case, researchers have developed single-item scales: They describe a trait to a respondent and ask them to place themselves on that scale (*e.g.*, five-item Big Five scales (Aronson, Reilly, & Lynn, 2006; Bernard, Walsh, & Mills, 2005; Gosling et al., 2003; Woods & Hampson, 2005)).

Experiments of demonstrating the correctness of the shortened scales typically involve administering both long and short forms to participants, and the developers show that their abbreviated instruments measure equivalent concepts as the long forms. However, Credé, Harms, Niehorster, and Gaye-Valentine (2012) have shown that longer personality scales tend to outperform the short forms, in that the longer forms are better able to explain the established personality dimensions. This shortcoming becomes a problem when researchers are looking for new personality dimensions to add: The short form overlooks some of the variance

among individuals that could be captured by the full scale, leading researchers to overestimate the effect of the new dimension.

2.2.3 Personalizing questionnaires: Adaptive testing

Personalizing a questionnaire allows past responses to influence future questions that will be chosen to be asked, so that a respondent answers the questions that are most relevant to them. Most work in this area has been done in the context of testing, where questions measure a respondent's skill or aptitude level (*e.g.*, (van der Linden & Glas, 2010)). However, the concept of adapting survey items administered to an individual can apply to any survey that seeks to measure some trait of the respondent through their answers to questions.

Aptitude or ability tests seek to estimate an examinee's achievement through their answers to test questions. Different questions are more useful for assessing test-takers of different skill levels (*e.g.*, extremely easy questions are not very helpful in assessing the achievement of a high-scoring individual, just as extremely difficult questions are not very helpful in assessing the achievement of a low-scoring individual). The goal of adaptive testing is to find the most relevant questions to ask an examinee to measure their achievement accurately, without making them answer too many questions. Typically, adaptive testing algorithms use item response theory (IRT) (Lord, 1980) to consider *individual* test questions through an item response function that expresses the probability of a correct answer by an individual at a particular skill level. The item response function has three parameters: the pseudo-chance score level (how easy it is to guess the correct answer), item difficulty (how hard it is to answer the question), and discriminating power (how much the skill level influences question response). In contrast, classical test theory assumes all items contribute equally to an assessment outcome, for all test-takers. According to Weiss (1982), an IRT-based adaptive testing framework has the following three components: (1) a way to choose the first item to ask, (2) a way to score items and choose the next item to ask during test administration, and (3) a way to choose to end the test, based on an individual's performance.

For example, Weiss and Kingsbury (1984) introduce adaptive mastery testing to assess a student's achievement level $\hat{\theta}$, specifically how the estimated achievement level compares to a "mastery level," θ_m. At each time point, a question is selected which gives the maximum information at the student's current estimated mastery level and asked. As the student answers questions, the estimate $\hat{\theta}$ is updated, along with a confidence interval. Once the confidence interval for $\hat{\theta}$ no longer includes θ_m, the test is finished and the student's mastery level is assigned as sufficient or not (depending if θ_m lies above or below the confidence interval for $\hat{\theta}$).

Computerized adaptive testing (CAT) approaches use a variety of item selection criteria during the testing process. One selection criterion popular in the early days of CAT was maximum Fisher information (*e.g.*, (Birnbaum, 1968)). At the time, computing capabilities were insufficient for Bayesian item selection, so Owen (1975) proposed using a normal approximation for the true posterior, for the selection criterion of minimizing the posterior variance of the estimate of ability θ. Later van der Linden (1998) performed Bayesian item selection using exact quantities for criteria of maximum posterior expected information, maximum predicted expected information, minimum predicted expected variance, and maximum predicted posterior expected information. Because the maximum Fisher information criterion performs best only when the predicted ability estimate $\hat{\theta}$ is near the true ability θ, alternate selection criteria attempted to generalize better across the range of possible θ values. Such criteria include the general weighted information criterion, which weights the information across all possible values of θ (Veerkamp & Berger, 1997); the interval information criterion, which chooses the maximum information in some confidence interval of θ (Veerkamp & Berger, 1997); and a global information measure that uses Kullback-Leibler item information to measure the distance between the current estimate and the true value (H.-H. Chang & Ying, 1996).

There has also been a significant body of research studying how to balance the statistical efficiency of an adaptive testing method (*i.e.*, getting a good estimator while administering a small number of items) with the practical constraints of an adaptive testing environment, such as item bank management. For example, a test that is administered to many people has its items refreshed regularly, to avoid issues of released questions or test-retakers influencing scores of subsequent test-takers. For many adaptive item selection criteria, early questions, before the estimate for a testee is accurate, will favor items with high discriminating parameters, for all testees. This means that these items are overexposed and will need to be refreshed frequently, while

other items are never or rarely asked. Approaches to diversify the initial questions of a CAT procedure include adding randomization to the item selection process (*e.g.*, (McBride & Martin, 1983; Sympson & Hetter, 1985; Kingsbury & Zara, 1989; Revuelta & Ponsoda, 1998; Chen, Ankenmann, & Spray, 1999; Barrada, Olea, Ponsoda, & Abad, 2008)), constraining eligible items to those that have not yet reached an exposure threshold (*e.g.*, (Revuelta & Ponsoda, 1998; Leung, Chang, & Hau, 2000)), and stratifying items by discrimination parameter so that items are selected from progressively more discriminating groups of questions (*e.g.*, (H.-H. Chang & Ying, 1999; H.-H. Chang, Qian, & Ying, 2001; H.-H. Chang & van der Linden, 2003; Cheng, Chang, & Yi, 2007; H.-H. Chang & Ying, 2008; Cheng, Chang, Douglas, & Guo, 2009)).

Adaptive tests have been shown to be as reliable and valid as conventional tests (with static question orders), while reducing test length up to 50% (Weiss, 1982; Weiss & Kingsbury, 1984). More recently, IRT-based adaptive testing has been used for diagnoses of mental health disorders through patient questionnaires (Gibbons, Weiss, Frank, & Kupfer, 2016) and measurements of political knowledge from public opinion surveys (Montgomery & Cutler, 2013). In both applications, the adaptive method resulted in equivalent or better outcomes with fewer questions.

These IRT-based approaches require training data to learn the item response parameters for each question. Most researchers recommend at least 1000 training samples to estimate these parameters (De Ayala, 2013). When the item response function accurately captures the structure of test questions, IRT-based adaptive testing approaches are good at modeling how questions inform estimates of respondents' knowledge. However, these IRT-based approaches do not consider question cost when selecting a question to ask—they consider only the impact that question will have on the understanding of that person's trait estimate.

2.2.4 Cost-effective personalized data collection

Combining the two goals of short form questionnaires (burden reduction) and adaptive testing (personalized data collection) yields cost-effective personalized data collection: determining which information is most relevant to collect from a particular individual, while taking into account the cost of acquiring that information. Recent past work has considered this problem, chiefly in a machine learning setting of test time feature acquisition: choosing which (costly) features to acquire to make a prediction on a new test point. Section 2.1 covers this area in more detail.

In summary, this prior work in questionnaire scoring has looked at scoring questionnaire responses (*e.g.*, (Digman, 1990)), reducing survey burden by creating short form questionnaires (*e.g.*, (Stanton, Sinar, Balzer, & Smith, 2002)), and personalizing tests (which can be considered a type of questionnaire) to a respondent's current score estimate (*e.g.*, (Weiss & Kingsbury, 1984)). However, there is a gap when it comes to cost-effective burden reduction for questionnaire scoring. This thesis fills that gap by personalizing questionnaire administration to individual respondents, while incorporating question-specific costs into the question selection rule.

2.3 Surveys

Unlike test-time feature acquisition or questionnaire scoring, the motivation of which is gathering the most relevant information for a prediction or score, surveys typically have the goal of collecting complete data from a population.

2.3.1 Respondent burden, respondent engagement, and data quality

Respondent burden in surveys is multi-faceted, with multiple factors influencing what is ultimately a subjective measure for the respondent. Some of these factors are survey length (*e.g.*, number of questions or estimated time), respondent effort, respondent stress, or frequency of interviews (Bradburn, 1978). A respondent's perception of the survey task has a significant direct effect on self-reports of survey burden (Fricker, Yan, & Tsai, 2014). Recent work has demonstrated that telling respondents that they have been screened into a longer or shorter survey, with a longer or shorter expected time commitment, can influence their perception of survey burden. Those who were told they were screened into a longer survey reported more burden than those who also received the longer survey but were not informed of their selection (Yu, Fricker, &

Kopp, 2015). Reported survey length is one factor that influences a respondent's decision to begin a survey: fewer people start surveys that are announced to take more time to complete (*e.g.*, (Walston, Lissitz, & Rudner, 2006; Marcus, Bosnjak, Lindner, Pilischenko, & Schütz, 2007; Galesic & Bosnjak, 2009)). However, this phenomenon is not universal (*e.g.*, (Cook, Heath, & Thompson, 2000)). Furthermore, response quality tends to decrease as the survey progresses (*e.g.*, (Galesic, 2006; Barge & Gehlbach, 2012)).

As survey response rates have been dropping (*e.g.*, (Porter, 2004; Shih & Fan, 2008)), researchers have begun looking at how to motivate respondents to fill out these surveys, to avoid having survey results not represent the full population. Commercial surveys often pay respondents, but compensation does not necessarily ensure thoughtful responses—participants still exhibit satisficing behavior in paid surveys (*e.g.*, (Barge & Gehlbach, 2012; Kapelner & Chandler, 2010)). Incentivizing respondents with something dependent on the quality of their answers, like a personalized prediction or calculation, can motivate them to provide correct data. For example, Angelovska and Mavrikiou (2013) design an online questionnaire that gives the respondent feedback on their level of procrastination, based on their responses. Their experiments find that the questionnaire that promises feedback has lower breakoff rates than the standard questionnaire in which the respondent does not receive personalized feedback. Marcus et al. (2007) find that offering personalized feedback increased response rates for low-salience surveys but had no significant effect for high-salience surveys. However, not all potential survey respondents react in the same way to changes in survey design. For example, meta-analyses of incentives and response rates in mail surveys are inconsistent in the effect of the amount of a prepaid incentive on response rates: Church (1993) found that greater incentive amounts yield greater increases in response rates, while Fox, Crask, and Kim (1988) found diminishing returns for increased incentive amounts. Leverage-saliency theory (Groves, Singer, & Corning, 2000; Groves, Presser, & Dipko, 2004) hypothesizes that different people respond to survey requests differently, based on what aspects of the survey they are interested in (*leverage*) and how much emphasis the interviewer or survey designers place on those aspects (*saliency*).

A promising venue for survey personalization to increase respondent motivation and lower breakoff is by using *paradata* (originally "process data" (M. P. Couper, 1998)) collected as the respondent answers an online survey (*e.g.*, time spent on page, mouse clicks (Kaczmirek, 2008)) to model user engagement (M. P. Couper et al., 2010). Past work in paradata analysis has focused on how paradata reveal respondents' survey-taking process (*e.g.*, (Heerwegh, 2003)), how paradata can identify usability issues in surveys (*e.g.*, (Healey, 2007)), and how paradata can inform population-level adaptive design (*e.g.*, (M. Couper & Wagner, 2012)). For example, Bassili and Fletcher (1991) and Heerwegh (2003) showed that participants who take longer to answer questions about attitudes also tend to change their minds when presented with counterarguments to their original response. They conclude that people with less attitude stability need more time to come up with an answer than people who are already confident in their attitudes (Heerwegh, 2003). Paradata about users' interactions can also reveal that a survey instrument is not designed well (*e.g.*, Healey (2007)). For example, drop-down menus in online surveys increased item response times (Healey, 2007). Question length (*e.g.*, (Bassili & Scott, 1996; Yan & Tourangeau, 2008)) and complexity (*e.g.*, (Wagner-Menghin, 2002; Yan & Tourangeau, 2008)) also increase response times. Paradata have been used for adaptive survey design too. For example, paradata have been used for predicting the likelihood that a unit will be interviewed on a next contact attempt (Groves & Heeringa, 2006; M. Couper & Wagner, 2012) and for analyzing differences in population estimates as contacts are made (Lundquist & Särndal, 2013).

Survey breakoff is influenced by respondent factors (*e.g.*, demographics), survey design (*e.g.*, topic), and page and question characteristics (*e.g.*, question type) (Peytchev, 2009). Respondent interest and burden are negatively and positively, respectively, associated with breakoff, and respondent and survey factors influence these components too (Galesic, 2006). Furthermore, lower-quality answers (*e.g.*, skipped items) often precede breakoff (Galesic, 2006). Survey action can be taken to increase user engagement and response rates.

2.3.2 Adaptive survey design

Adaptive survey design (ASD) attempts to improve survey quality (often in terms of achieving a higher response rate or lower error) by giving respondents custom survey designs, rather than the same one (Schouten, Calinescu, & Luiten, 2013; Wagner, 2008). Usually ASD tries to minimize nonresponse, and designs involve factors like number of follow-ups, which can be costly. The general technique is to maximize survey quality while keeping costs below a budget. For example, Beaumont, Bocci, and Haziza (2014) adaptively choose

how many contact attempts to make for a household sampled for a telephone survey by minimizing the variance of the estimator to determine how much effort to expend (*i.e.*, number of contact attempts) on each household.

Dynamically ordering questions in a survey, based on answers given to previous questions, can be considered a type of adaptive survey design. Researchers have chosen a variety of aspects of survey design to adapt with ASD (see (Tourangeau, Michael Brick, Lohr, & Li, 2017) for a review), including adaptive contact strategies to schedule call attempts at the time a respondent is most likely to answer (Wagner, 2012), case prioritization to exert more effort to attain responses from certain households (*e.g.*, (Beaumont et al., 2014; Wagner et al., 2012)), and stopping rules to determine when to stop collecting information from groups of respondents whose overrepresentation might bias population-level estimates (*e.g.*, (Lundquist & Särndal, 2013; Särndal & Lundquist, 2014; Särndal & Lundquist, 2014)).

Often in ASD, changes in survey design happen between *phases* of the survey, where the exact same survey protocol (*e.g.*, sampling frame, survey mode, measurement conditions) is in place within a phase and results from that phase inform changes to the protocol for the next phase. Groves and Heeringa (2006) introduce an approach they call responsive survey design, which uses indicators of the cost and error of design features to make decisions about how to change the survey design in future phases and then combines data from all phases into a final estimator. They also introduce the concept of *phase capacity*—once a stable estimate has been reached in a design phase, it is unlikely that expending more effort in that phase will result in a better estimate. Their definition of "effort" focuses on collecting participants for each phase. They propose the use of error-sensitive indicators to identify when a phase has reached capacity and no more participants need to be recruited for that phase. This notion of phase capacity could extend to reaching a stable estimate of a participant's survey-answering, and no more questions need to be asked.

Previous approaches to ASD that are most related to the work in this thesis are those that choose respondents to prioritize. Such case prioritization approaches can be considered as choosing to obtain information, in the unit of a respondent, taking previously acquired information (previous responses or paradata that have been collected during the survey-taking process) into account. Correspondingly, the work of this thesis chooses to obtain information, but in the unit of individual answers from respondents, also taking previously acquired information (answers that have already been given or paradata) into account.

2.4 Summary

The domain-specific approaches reviewed in this section adapt data collection to the concerns of their field—machine learning techniques for test-time feature acquisition trade off increases to prediction quality against feature cost, adaptive testing methods choose test items to administer based on information at the testee's current skill level (with no question-specific measure of cost), and adaptive survey design approaches choose changes to make to survey design to improve response rates or representativeness. In the next chapter, we present a general framework for adaptive data collection that can encompass many of these prior works, by defining the appropriate utility and cost metrics to quantify how useful and how expensive an item is to collect, in the context of the information that has already been collected.

Chapter 3

A general framework for question ordering

Our goal is to personalize question ordering for an individual providing information, depending on our current knowledge of them. This knowledge could come from answers they have given to previous questions, as well as information collected through paradata during the information-providing process. Because we do not know ahead of time the budget for question asking, we present a greedy approach for iterative question selection: at each time step, choose the question that optimizes a selection criterion; ask that question and receive an answer; update the current knowledge base; and continue the process until all questions have been asked, the allowable budget for asking questions is exhausted, or the respondent stops providing answers.

3.1 A criterion for iterative question selection

The question selection rule trades off the expected utility of having an answer to a question (or a set of questions) with the cost of getting the answer:

$$q^\star = \arg\min_q \left(-\mathbb{E}[U(q)] + \lambda c_q\right), \tag{3.1}$$

where q is a question or set of questions, $U(q)$ is the utility of q, c_q is the cost of q (which may be context-dependent and therefore not known until evaluation time), and λ is a tradeoff parameter. We want to maximize utility while minimizing costs. Figure 3.1 summarizes how this iterative question selection process works for a sample at test time. Definitions of question utility and cost will vary according to the application and the purpose of the data collection; here we present an overview of possible interpretations for utility and cost, which will be expanded for example applications in our experiments in later sections.

This method can be generalized to order *modules* of related questions, rather than individual questions. Reasons to present questions in modules rather than purely sequentially include (1) presenting related items in a group can reduce the *cognitive burden* required of a respondent to answer the group (*e.g.*, if a set of questions asks the respondent about various aspects of their commute, as the American Community Survey does, it will be easier for the respondent to answer those commute-related questions as a unit rather than scattered throughout the entire questionnaire) (Tourangeau, 1984) and (2) imposing a standard order on certain questions that are susceptible to *order effects* (Sudman, Bradburn, & Schwarz, 1996) can ensure that all participants understand and answer questions in the same way, even when question order is determined dynamically.

3.1.1 Question utility

Intuitively, question utility captures how "useful" a candidate question is. The exact definition for utility depends on the application (*i.e.*, what do we value as useful?) and the data (*i.e.*, how can we calculate this value?). For example, if the application is a prediction task, we can use the impact a question will

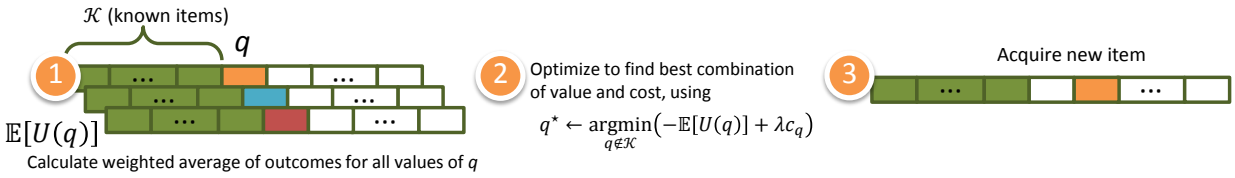


Figure 3.1: Our method iteratively increases the set of known items. **Step 1:** Given a set of known items, it first calculates the expected utility and cost of acquiring each unknown item. This is repeated for all unknown items. **Step 2:** It optimizes for the best combination of utility (as calculated in Steps 1 and 2) and cost. **Step 3:** The best next item is acquired. The process can be repeated until the all questions have been asked or the budget for information collection has been exhausted.

have on the prediction quality as a definition for utility. One aspect of prediction quality is *certainty*. The calculation of prediction certainty as a measure of utility will depend on the data and the predictive model being used: if the value to be predicted is continuous, one measure for prediction certainty is the width of the prediction interval, where a wider interval means a less certain prediction. The mathematical definition for the prediction interval width will depend on the predictive model being used. If the value to be predicted is discrete, one measure for prediction certainty is the distance of the sample from the decision boundary. Again, calculation for this measure of certainty (the distance from the decision boundary) will depend on the predictive model.

To be dynamic, these definitions of utility will take into account information already known about the sample. Such information may come from previously answered questions or paradata collected in the survey process. Paradata in particular will be helpful for informing a utility function that captures respondent engagement.

When defining a metric for question utility, it can be challenging to reflect the multiple purposes of a survey in a single utility function. Large-scale government surveys in particular often have multiple federal agencies who care about different questions. Developing a composite utility function that meets the needs of all stakeholders in a multi-purpose survey is important, but also complicated and context-specific, and a topic for continued work. This thesis illustrates examples of the following utility functions: prediction certainty (for continuous outputs, in Sections 4.4.4 and 4.4.5, and discrete outputs in Section 4.4.6), total response variation (Section 5.6), information gain (Sections 6.3 and 6.2), and response likelihood (*i.e.*, the probability that a respondent will not break off from the survey; Section 6.3).

3.1.2 Question cost

The cost of a question reflects how "difficult" it is to get an answer for that question. Different applications have different cost measures that are best suited to them. Examples of cost include (1) the amount of resources needed to answer the question (*e.g.*, time, money, battery, effort), (2) the likelihood that the question will not be answered (*i.e.*, item nonresponse rate), (3) the likelihood that the question will cause the respondent to stop the survey (*i.e.*, item breakoff rate), and (4) a combination of multiple types of cost. These costs can be predefined according to rules or determined empirically from collected data. Respondent burden in surveys is multi-faceted, with multiple factors influencing what is ultimately a subjective measure on the respondent (*e.g.*, (Bradburn, 1978; Fricker et al., 2014; Yu et al., 2015)). Thus, it is difficult to measure the actual burden a question poses to a respondent. Our illustrative applications in Chapters 4, 5, and 6 use simple proxies for respondent burden. However, this framework is general, and so any more cognitively precise measure can be used, were it available.

When determining question order is deferred to test time, these costs can be *context-dependent*. That is, the current situation of data collection can inform the costs of future questions that might be asked. For example, in a system that makes medical diagnoses and can request tests (*e.g.*, (Ferrucci et al., 2013)), it will be less costly to request an invasive biopsy if a surgery is already scheduled in that area. Context-dependent

costs can also be used to address questions that are sensitive to order effects (McFarland, 1981). At response time, the cost of a dependent question can be assigned to be infinite when its prerequisite question(s) have not already been asked. Thus, such a question would never minimize the objective in Equation 3.1 and would never be chosen if the respondent has not already seen questions that need to precede it.

This thesis illustrates examples of the following costs: effort required to answer questions (Section 4.4.4), battery drain of collecting mobile data (Section 4.4.5), time to compute features (Section 4.4.6), item response times (Sections 5.6 and 6.3), and item nonresponse rates (Section 6.2).

3.2 Terminating data collection

A key component of computerized adaptive testing procedures is a criterion for determining when to stop administering test items (Weiss, 1982). Typically, the test is terminated once the estimate about the testee is certain enough (*e.g.*, (Weiss & Kingsbury, 1984; Kamakura & Balasubramanian, 1989)). A similar stopping criterion could be used for prediction-guided DQO, but most often surveys aim for complete response. In that case, it is advantageous to ask questions as long as the respondent is willing to answer them. One exception could be for longitudinal surveys, where respondent attrition is partly due to panel fatigue (Laurie & Scott, 1999). Terminating a panel survey before completion would reduce respondent burden and could increase willingness to participate in future surveys.

3.3 Assumptions and requirements

Most generally, this approach requires only definitions of question utility and cost and ways to calculate the expected utility of each candidate question given what is known. In practice, a training set of question responses is needed: measures of utility often depend on statistical properties of samples and estimators, and distributions of question responses need to be known to calculate expected values.

3.4 Relationship to prior work

The generality of this framework means that much of the prior work in adaptive data collection presented in Chapter 2 can be expressed in the DQO framework. Given appropriate definitions for utility and cost, any method that gathers information in stages has a representation in DQO. For example, all of the computerized adaptive testing (CAT) methods in Section 2.2.3 can be expressed as DQO variants by using whatever item selection criterion the CAT algorithm (simple functions like maximum Fisher information (*e.g.*, (Birnbaum, 1968)), maximum global information (*e.g.*, (H.-H. Chang & Ying, 1996)), minimum variance (*e.g.*, (van der Linden, 1998)), *etc.*; or more complex rules incorporating randomization (*e.g.*, (Barrada et al., 2008)), item stratification (*e.g.*, (H.-H. Chang & Ying, 1999)), *etc.*) uses as the utility metric. None of the CAT approaches considered question-specific costs, which can be translated into DQO by setting costs for all items to zero.

Categories of methods discussed in Chapter 2 that cannot be subsumed as special cases of the DQO framework include those that do not sequentially gather information and use that information for subsequent item requests. For example, adaptive contact strategies that optimally schedule call attempts to improve response probability (*e.g.*, (Wagner et al., 2012)) do not fit into the DQO framework, since there is no incremental collection of data that inform later pieces of information to gather. Similarly, the test-time feature acquisition approaches that essentially learn rules for feature order at training time (the sequence-learning approach (*e.g.*, (Strubell et al., 2015)) or the Markov decision process approach (*e.g.*, (He et al., 2012)) cannot be described using the DQO framework since the mechanism for acquiring information is determined at training time and is not updated as data are gathered at collection time.

Table 3.1 summarizes if and how the previous work covered in Chapter 2 can be represented in the DQO framework. When decisions about data collection are made at collection time, it is often possible to define utility and cost functions to reproduce the method with DQO. One exception is the work on using adaptive survey design to decide when to stop data collection (Lundquist & Särndal, 2013; Särndal & Lundquist, 2014; Särndal & Lundquist, 2014)—although the decision of when to continue or stop gathering survey responses is

technically made at data collection time, the procedure for choosing the stop time is fixed prior to collection time (*e.g.*, ending data collection once the response rate in a subgroup reaches a threshold). Thus, as data are collected, response rates (or other indicators) are monitored until a prespecified condition is met and data collection is terminated.

Table 3.1: If and how previous work can fit into the dynamic question ordering framework. When decisions about data collection are made at collection time, it is often possible to define utility and cost functions to reproduce the method with DQO.

Category	**Paper**	**Decision being made**	**Decision time**	**Utility**	**Cost**
Test-time feature acquisition	(He et al., 2012)	Feature to acquire	Pre-collection		
	(He et al., 2013)		Pre-collection		
	(Weiss & Taskar, 2013)		Pre-collection		
	(Xu et al., 2014)		Pre-collection		
	(Samadi et al., 2015)		Pre-collection		
	(Shi et al., 2015)		Pre-collection		
	(Strubell et al., 2015)		Pre-collection		
	(Pattuk et al., 2015)		Collection	Prediction uncertainty	Privacy
Computerized adaptive testing	(Birnbaum, 1968)	Question to ask	Collection	Maximum Fisher information	0
	(Weiss & Kingsbury, 1984)			Max info, min posterior variance	
	(H.-H. Chang & Ying, 1996)			Max global info	
	(van der Linden, 1998)			Max posterior expected info	
	(McBride & Martin, 1983)			Max info + random	
	(H.-H. Chang & Ying, 1999)			Stratified max info	
	(Leung et al., 2000)			Content-restricted max info	
Adaptive survey design	(Groves & Heeringa, 2006)	Protocol to use	Collection	Phase capacity	Survey effort (*e.g.*, time)
	(Wagner et al., 2012)	Cases to prioritize	Pre-collection		
	(Wagner, 2012)	Contact schedule	Pre-collection		
	(Schouten et al., 2013)	Interviewer assignment	Pre-collection		
	(Lundquist & Särndal, 2013)	Stopping collection	Collection		
	(Särndal & Lundquist, 2014)	Stopping collection	Collection		
	(Särndal & Lundquist, 2014)	Stopping collection	Collection		
	(Beaumont et al., 2014)	Contact schedule	Pre-collection		

Chapter 4

Prediction-guided question ordering

Here we consider the scenario in which a user is providing information to receive a personalized prediction based on the information they provide. We assume a predictive model has already been developed from a training set, and we want to make a prediction on a new test point. In this setting, it is likely that not all features will be needed to make a reasonable prediction. Therefore, we want to ask the most cost-effective set of questions that will maximize prediction quality while minimizing collection costs. Additionally, in this scenario, users receive a direct benefit (*i.e.*, the personalized prediction) and are inherently motivated to provide information. On the other hand, feedback in the form of sequential predictions can influence their decision to continue providing information, if they feel that they have received a useful enough prediction.

When features are being collected to refine a prediction, a natural utility metric to use for test-time feature selection (Equation 3.1) is one that reflects how an additional feature f contributes to the *quality* of the subsequent prediction. An ideal solution would accommodate a variety of prediction algorithms, allow for feature-specific and context-dependent costs, and support multiple options for prediction utility when requesting information to refine a prediction. Delaying the order determination until test time (rather than learning it from training data) is a requirement for context-dependent costs (since they are not known until test time).

We developed an approach to test-time Feature Ordering with Cost and Uncertainty Score (FOCUS) that has all of these key properties. A basic assumption of our approach is that it is being applied in the context of supervised machine learning. In addition, we assume that feature cost is only an issue at test (or more generally deployment) time—at training time, the complete set of features associated with each label is assumed to be available. Finally, although not required, our approach benefits from a prediction algorithm that is robust to making predictions when not all features are available (*predictions on partial information*).

FOCUS is an instantiation of the question ordering framework from Chapter 3, where the utility of a feature is defined to be the *certainty* of the subsequent prediction made using all known features plus the newly acquired feature. We chose to optimize for prediction certainty due to past work demonstrating the importance of giving people estimates of certainty along with predictions (*e.g.*, (Hirschberg et al., 2011)). Prediction certainty is available in most prediction algorithms and reflects how likely the true value is to coincide with the estimated value. For example, in regression, a prediction interval width indicates the range of values the true value is likely to fall within (Weisberg, 2014). A narrower prediction interval corresponds to a more certain prediction. For classification, certainty can often be formulated as distance from the decision boundary. However, certainty can be easily replaced with another utility metric (*e.g.*, prediction error) in the question ordering framework.

As shown in Figure 3.1, FOCUS operates iteratively at prediction time. Given an instance with known (dark green) and unknown (white) features, it first calculates the expected value of a feature (Step 1) for all features that are not known. Next it calculates the best next feature by optimizing a function that combines the prediction value of each feature with its cost (Step 2). The new feature is acquired (Step 3) and the process repeats. At any iteration, a prediction can be made.

Whether or not the prediction at each step is shown to the user will depend on the application. For example, if the application is gathering information from sensors with no user input, it does not make sense to display the sequential predictions to the user. In this case the algorithm may stop when costs become

too high, all features are acquired, or accuracy achieves a certain threshold. This can be determined on an application-by-application basis. On the other hand, if the user is answering questions to get a personalized prediction, then showing them the partial predictions can keep them engaged in answering questions and help them to decide if continuing to answer questions is valuable to their goals.

4.1 The FOCUS algorithm

FOCUS assumes that a model trained on all feature values is available and that, for a new test point, we want to provide the best (and lowest cost) prediction possible, given that feature values are costly to acquire. As shown in Figure 3.1, FOCUS sequentially estimates the value of each feature (Step 1) and selects a next feature to ask (Steps 2 and 3).

4.1.1 Calculating the utility of a feature f

In Step 1 of FOCUS, the expected utility of a feature is calculated. Since the true next prediction uncertainty depends on the actual value for the next feature that is acquired, we cannot directly calculate $\mathbb{E}[U(f)]$. However, we can break this down into two parts.

Calculating $U(f)$, the prediction utility for a specific possible value of feature f, depends on the exact nature of the prediction problem. $U(f)$ takes as input a feature vector containing the known features plus a hypothetical value r for feature f. Typically, utility is calculated by making a partial prediction using those values and estimating prediction accuracy, uncertainty, *etc*

This is repeated for all values r that are in the range of potential values R for f (Step 1 of Figure 3.1). If a feature is continuous, we pick bins appropriate for the values that appear in the training set, and then the midpoint for each bin for feature f is used as the set of values f can take on.

There have been several approaches to making predictions under partial information, and FOCUS is compatible with any of them. Reduced-feature models use only features whose values are known in making predictions. These models may be calculated at training time (*e.g.*, (Xu et al., 2014)) or dynamically constructed at test time (*e.g.*, (Friedman, Kohavi, & Yun, 1996)). Another option is to impute missing values and use the full-feature model on the combination of known and estimated features to make a prediction. Hybrid approaches combine reduced-feature modeling with imputation (*e.g.*, (Saar-Tsechansky & Provost, 2007)). However, in the end, FOCUS is agnostic about how predictions are made and how U is defined.

4.1.2 Calculating the expected prediction utility of a feature f

Given a way to calculate $U(f)$, $\mathbb{E}[U(f)]$ can easily be defined. We calculate the expected utility of a prediction that includes feature f by taking a weighted average of the utility calculated for each possible value of f:

$$\mathbb{E}[U(f)] = \sum_{r \in R} p(z_f = r) U(z_{f:=r}), \tag{4.1}$$

where $p(z_f = r)$, the probability that the f-th feature's value is r, is calculated empirically from the training set, and the notation $z_{f:=r}$ means that the f-th component of feature vector z is replaced with the value r. Algorithm 2 in Appendix A summarizes this process in pseudocode.

This process is then repeated for all unknown features.

4.1.3 Optimizing for the best next feature

In its middle step (for each iteration), FOCUS optimizes for utility of the next feature, penalized by the cost of that feature. Our selection rule, illustrated in Step 2 of Figure 3.1, trades off the expected utility of the next prediction, for each candidate feature, with the cost of that feature:

$$f^\star = \arg\min_{f \notin \mathcal{K}} \left(-\mathbb{E}[U(f)] + \lambda c_f \right), \tag{4.2}$$

where $\mathbb{E}[U(f)]$ is the expected utility calculated in Step 1 of FOCUS, and λ controls how much weight we give to the cost c for each feature. Algorithm 3 summarizes this process in pseudocode.

As with utility, cost is calculated in a problem-specific fashion, based on the feature vector of currently known features. This allows the cost function to consider context-dependent information such as the values of other features that are already known. Cost could be measured in time necessary to acquire a feature, either computationally or due to dependencies; direct impact on the user such as interrupting her to ask a question; or indirect impact on the user such as drawing down the battery life of her phone.

4.2 Regression: Predicting continuous values

Our first two applications (predicting household energy consumption and predicting student stress levels) involve predicting continuous values, so we use prediction interval width to measure uncertainty: a narrower prediction interval corresponds to a more certain prediction. Because we are imputing missing features to make predictions with partial information, we use the measurement error model (MEM) (Fuller, 2009) to capture error associated with estimated features. Unlike traditional regression models, MEMs do not assume we observe each component x_f exactly—there is an error δ_f associated with the estimation:

$$z_f = x_f + \delta_f, \text{ where } \mathbb{E}[\delta_f | x_f] = 0.$$

Prediction $\hat{y}$ still depends on the *true, unobserved* value x:

$$\hat{y} = \hat{\beta}^T \bar{x} = \hat{\beta}(\bar{z} - \bar{\delta}),$$

where $\hat{\beta} \in \mathbb{R}^{d+1}$ is the parameter vector learned on the training set X (recall that all feature values are known at training time). The notation $\bar{x}, \bar{z}, \bar{\delta}$ means vectors x, z have a 1 appended to them and δ a 0 to account for the constant term in the regression. Let $\bar{X}$ extend this notion to the training matrix: $\bar{X} = [\mathbf{1}^n X]$.

We can calculate a $100(1-\alpha)\%$ prediction interval for a new point z as

$$\hat{y} \pm t_{n-d-1;\alpha/2} \sqrt{\hat{\sigma}^2 \left(1 + \bar{z}^T (\bar{X}^T \bar{X})^{-1} \bar{z} + \bar{\delta}^T (\bar{X}^T \bar{X})^{-1} \bar{\delta}\right)}, \tag{4.3}$$

where the $\bar{\delta}^T (\bar{X}^T \bar{X})^{-1} \bar{\delta}$ term accounts for error from estimated features and $t_{n-d-1;\alpha/2}$ is the value at which a Student's t distribution with $n-d-1$ degrees of freedom has cumulative distribution function value $\alpha/2$. We can estimate δ from training data by calculating the error of predicting each feature with k-NN, from the other features. We also estimate $\hat{\sigma}^2$, the regression variance, from training data.

When using prediction certainty as the measure of feature utility (where feature utility means how much a feature will influence the prediction quality), a more certain prediction (*i.e.*, a more useful feature) will have a smaller prediction interval width than a less certain prediction. Thus, in this case, we want to minimize prediction interval width (equivalent to minimizing uncertainty, or maximizing certainty/utility), so we use the negative prediction interval width as the utility measure for Equation 3.1, which then reduces to

$$f^\star = \arg\min_f \left(\mathbb{E}[W(f)] + \lambda c_f\right), \tag{4.4}$$

where $\mathbb{E}[W(f)]$ is the expected prediction interval width of the next prediction that includes feature f.

4.3 Classification: Predicting discrete values

The utility metric used to measure prediction uncertainty for regression, the prediction interval width, is not applicable for classification. However, we can define a measure of prediction uncertainty that is appropriate for classification. For classification, the probability that a test instance belongs to a particular class is the certainty of its prediction—a lower class probability—means a less certain prediction. More generally, class probability can be seen as the distance from the decision boundary—when a point is farther from the decision boundary, the prediction is more certain.

4.4 Applications and experiments

To demonstrate the usefulness of FOCUS for cost-effective interactive predictions, we implemented it for several prediction algorithms and applications. First, we consider the case of providing personalized energy estimates for prospective tenants, where feature cost reflects how much effort a user must exert to provide answers about their energy-consuming habits and their new potential home. Next, we use FOCUS to make momentary predictions of stress in college students, where feature cost includes both battery drain from turning on mobile sensors in addition to the cost of interrupting the user to ask for feature values. In this example, we also consider *context-dependent* costs, in particular, the fact that the cost turning on a sensor at the expense of draining the battery is no longer an issue when the phone is charging. Finally, we apply FOCUS to the classification problem of identifying devices in photos to support opportunistic interactions with low user burden.

4.4.1 Methods

For each of these applications, we compare the quality and cost of successive predictions obtained with our method, FOCUS, with a variety of cost penalties λ, to that of a fixed-order baseline, which we call *Fixed Selection*. This baseline acquires features in the order of forward selection on the training data (resulting in an identical ordering for all samples). Forward selection is a greedy approximation to feature selection, which starts with an empty feature set and iteratively adds the feature that minimizes error to the predictive model (Harrell, 2001; Tropp, 2004).

In this problem, we make predictions when there are features that are still unknown (*i.e.*, not yet supplied by the user). In all three applications, we use imputation to estimate values for these unknown features before making a prediction at that stage in data collection. To make these imputations, we use k-nearest neighbors (k-NN) (Cover & Hart, 1967) as follows: First, we find the k points in the training set that are nearest to the test point, restricted to the features that are already known. Next, we estimate the value for each unknown feature in the test point as the mean or mode (depending whether the feature is continuous or discrete) of that feature in the k-nearest neighbors (see Algorithm 1 in Appendix A). Finally, we use this complete feature vector, which contains both known and imputed values, to make a prediction at the current step.

We simulate progressive addition of features by hiding values for "unknown" features (*i.e.*, pretending their values are truly unknown since we have not yet asked for and acquired their values), using FOCUS to choose a feature to acquire at each step, and unveiling that feature's value once it has been "asked" and "answered."

4.4.2 Performance metrics

Our validation measures performance at each stage of feature acquisition (*i.e.*, number of collected features) in terms of prediction certainty, error, and cost. Certainty and cost are defined as part of the FOCUS optimization objective (Equation 4.2).

Prediction certainty

Depending on whether the prediction is continuous or discrete, we use prediction interval width (Section 4.2) or predicted class probability (Section 4.3) to measure the certainty of a prediction made at each stage of feature acquisition.

Prediction cost

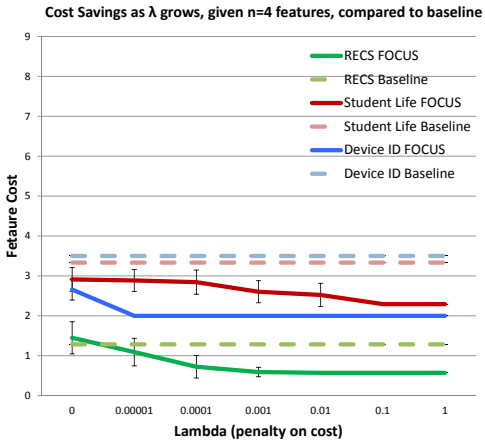
Prediction cost is defined as the total cost of all features acquired up to each point in feature acquisition. Cost is application-dependent, and the applications we describe here use costs related to user burden, battery drain, and computation time. The specific cost metrics used for each application are discussed in more detail in the following sections.

Prediction error

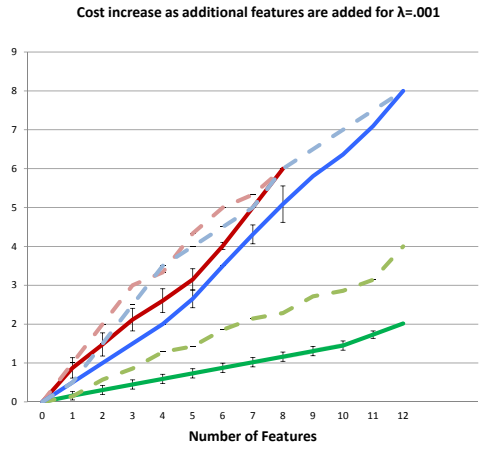
For error, we use mean absolute error for regression and zero-one loss for classification (*i.e.*, a sample incurs an error of 0 if its predicted value matches the true value and 1 otherwise) to compare our successive predictions to the true values. This is a conservative metric for error since it compares only a single predicted value to the true value (rather than taking into account the uncertainty associated with the prediction). For example, this metric will incur error when the true value is not the exact midpoint of a prediction interval, even when the true value does lie within the prediction interval.

4.4.3 Overview of results

Figure 4.1 illustrates the cost savings of FOCUS (solid lines) over the baseline (dashed lines) when various numbers of additional features have been provided for each of our three validation applications. The left plot (Figure 4.1a) shows that, as the cost penalty λ increases, the cost savings of FOCUS over the baseline also increases. As expected, increasing the cost tradeoff parameter λ favors asking inexpensive features near the beginning of the test-time feature acquisition process. The right plot (Figure 4.1b) shows that, for a fixed λ and with increasing numbers of additional features, FOCUS also maintains its cost advantage over the baselines, for all three applications.



(a) As the cost penalty λ increases, the cost savings from FOCUS increase, without significant loss of accuracy, compared to the fixed order baseline. On this plot, the number of features n is fixed at 4.



(b) FOCUS ensures that cost increases more slowly than the baseline, with $\lambda = .001$. The Student Life and Device ID sets converge because there are only 8 and 12 total features. RECS converges at 30 features (not shown).

Figure 4.1: Charts showing impact on cost (y axis) of λ and number of features, for FOCUS (solid lines) and the Fixed Selection baseline (dashed lines).

To summarize the trajectories of prediction cost, uncertainty, and error, we calculate areas under the curve for each of the metrics as each feature is added. Smaller values are better because they mean the algorithm spent less time in high cost, uncertainty, and error. Table 4.1 lists these values for FOCUS with seven different values of cost penalty λ, ranging from zero to one, and for the baseline. As expected, due to the two terms in the selection rule (uncertainty and cost), cost decreases as the penalty on cost increases, and uncertainty tends to increase as the cost penalty increases. There is no pattern in how error changes as λ increases. The baseline (Fixed Selection) error is often lower than FOCUS. This result is not surprising, due to the error-minimizing criterion of forward selection. FOCUS with a nonzero λ always has significantly lower cost than the baseline. Prediction uncertainty is sometimes lower with FOCUS than baseline (particularly for low values of λ). The metric that suffers the most with FOCUS, relative to the baseline, is error because it was not included in the optimization.

The following subsections discuss the specific problems, data, and results for each application in more detail.

Table 4.1: Areas under the curve for the cost, uncertainty, and error metrics from **FOCUS**, with a variety of cost penalties λ, and the **Baseline** (*Fixed Selection*): smaller values mean the algorithm spent less time in high uncertainty, error, and cost. Ideally, FOCUS will have lower cost and similar uncertainty as Baseline. This is shown in black. Numbers marked with $\star$ show where FOCUS was significantly lower than baseline at the $\alpha = 0.05$ level. Numbers in red, marked with †, indicate where FOCUS did significantly worse than baseline (higher cost, uncertainty, or error) at the $\alpha = 0.05$ level.

Method		RECS			StudentLife			Device ID		
		Cost	*Uncert.*	*Error*	*Cost*	*Uncert.*	*Error*	*Cost*	*Uncert.*	*Error*
FOCUS, $\lambda =$	0	1340.25†	41.376⋆	12.457	71.182⋆	22.897	5.770	94.788⋆	1.935	4.815
	.00001	1175.42⋆	41.375⋆	12.444	70.924⋆	22.898	5.769	87.647⋆	1.939†	5.241
	.0001	936.338⋆	41.371⋆	12.649†	70.030⋆	22.898	5.771	87.601⋆	1.940†	5.259
	.001	891.130⋆	41.430⋆	12.578†	66.704⋆	22.899	5.775†	87.638⋆	1.936	5.296†
	.01	885.000⋆	41.446	12.704†	63.303⋆	22.929	5.795†	82.001⋆	1.936	5.463†
	.1	885.000⋆	41.446	12.704†	61.303⋆	22.958	5.800†	80.547⋆	1.936	5.574†
	1	885.000⋆	41.446	12.704†	61.303⋆	22.958	5.800†	80.000⋆	1.926	5.333†
Baseline		1250.000	41.447	11.722	81.000	22.898	5.705	105.000	1.928	4.796

4.4.4 Predicting energy usage for prospective tenants

Selecting energy-efficient homes is important for renters, because in many climates energy costs can be a significant burden, and the choice of infrastructure influences energy consumption far more than in-home behavior (Dietz, Gardner, Gilligan, Stern, & Vandenbergh, 2009). However, there is a paucity of information available about expected energy costs pre-lease signing. Calling a utility to ask about prior costs may give incomplete or misleading information (since occupant behavior can influence energy usage as much as 100% (Seryak & Kissock, 2003)), and a carbon calculator typically requires users to answer (prohibitively) many questions that may require considerable research to answer. We can lower user burden by (1) learning the relationship between household features (home infrastructure and occupant behavior) and energy use from established datasets, such as the Residential Energy Consumption Survey (RECS) (U.S. Energy Information Administration, 2009), and then (2) using FOCUS to strategically select which instance-specific features are needed to make a confident prediction for a new household.

Background

Bottom-up methods for modeling residential energy consumption use features of individual households (Swan & Ugursal, 2009). Household features may include macroeconomic indicators, as well as occupant-specific features (*e.g.*, (Douthitt, 1989; Kaza, 2010)). For example, Douthitt (1989) examines fuel consumption for space heating in Canada, using household-specific values for occupant demographics, the housing structure, and fuel cost. Kaza (2010) uses the Residential Energy Consumption Survey (RECS) to estimate energy usage from low-level housing characteristics, with a quantile regression approach to separate the effects of variables on homes with different patterns of energy consumption.

The RECS dataset is our focus as well. RECS consists of data from 12,000 households across the United States, with energy consumption by fuel type (*e.g.*, electricity, natural gas) and around 500 features of each home and its occupants. We restrict our use of RECS to electricity prediction in a single climate zone where energy consumption is variable due to cold weather, giving us a subset of 2470 homes.

Cost metric

In Early et al. (2017), we assign cost using a three point scale: some features are free (available in rental advertisements), while the remainder are either low cost (*e.g.*, number of occupants) or high cost (*e.g.*, age of heating equipment). Free features are included in all predictions since they have no cost. Here, we use a more nuanced eight-point feature cost scale based on difficulty of acquiring a feature. Zero-cost features are "free" (*i.e.*, extractable from a rental listing); 1-2 are occupant-related; and 3-7 are unit-related (may require

a site visit and/or research). Table 4.2 lists the cost categories and an example feature in each. Table B.1 in Appendix B lists the cost categories for all features in the RECS 2009 dataset. Table 4.3 justifies our choice of "free" features (*i.e.*, features extractable from rental advertisements) by showing how frequently those features appear in the database of Rent Jungle, an online rental search platform[1].

Table 4.2: Feature cost reflects how difficult it is to acquire a value for a feature. Note that the cost of a feature might vary by home: for example, if a listing included pictures of the kitchen, a user could see the door arrangement of the refrigerator from the listing (cost 3). If there were no pictures, then it would be an easily visible feature (cost 5).

Cost	**Method of answering**	**Example feature from RECS**
0	Extractable from listing	Number of bedrooms
1	User probably already knows	If someone stays home during the day
2	Might have to check current home	Size of TV
3	Could find in rental listing, or call for easy answer	Washing machine in home
4	Look up online	Year housing unit was built
5	Easily visible during visit	High ceilings
6	Requires more effort to find out during visit	Type of glass in windows
7	Requires visit + looking something up	Age of heating equipment

Table 4.3: The features we define as extractable (*i.e.*, "free") appear in most of the listings on Rent Jungle. Geographic features associated with the city, zip code, or state include climate zone and whether the area is urban or rural, among others.

Feature	**Presence in Rent Jungle**
Number of bedrooms	85%
Number of full bathrooms	57%
Studio apartment	85%
City or zip code	99%
State	100%

Experimental setup

We first divide RECS into training (90%) and testing (10%) sets. At training time, we use forward selection (Tropp, 2004) on a randomly selected subset of 20% of the training data to choose 30 higher-cost features to add to the free features for prediction. We use ten-fold cross validation on the remainder of the training data to learn regression weights.

At test time, the goal is to make a prediction on a new test point and acquire costly features as needed to improve the prediction. We use FOCUS on each test point to obtain features in the order that optimizes the FOCUS objective (Equation 4.2).

Results

The regression model using FOCUS had significantly lower costs than the baseline for all values of λ except 0 with equivalent or better uncertainty (Table 4.1). For $\lambda = .001$, FOCUS yielded an average of 45% savings in cost compared to the baseline as features were added to the prediction. As expected, FOCUS performed worse in terms of error than the baseline did, since FOCUS optimizes prediction uncertainty, rather than error, when determining a test-specific feature order. In contrast, the baseline was chosen by optimizing

[1] http://www.rentjungle.com

prediction error on the training set. However, for this application in particular, it is more important that predicted energy costs fall inside the window of uncertainty (which the definition of a prediction interval ensures) than that they have an accurate dollar value (Hirschberg et al., 2011).

Conclusion

These results for energy prediction using RECS illustrate that the FOCUS method for adaptively selecting features to acquire at test time can provide predictions of similar quality (in terms of certainty and accuracy) at up to a 45% lower cost than a fixed order approach.

This application of predicting energy usage could assist prospective tenants in their decision making while searching for a rental unit. A system that implements FOCUS for energy usage predictions could collect information about a renter's energy habits and the infrastructure of rental units they are considering. The experiments using FOCUS in this application assumed that a user would be able to answer the next question asked. However, in a real-world setting such as an energy prediction system for prospective tenants, a user might not know the value of a feature. For example, a prospective tenant might be interested in getting an energy usage estimate for a rental unit before visiting the unit. In this case, they could provide information about their energy consumption habits (*e.g.*, preferred indoor temperature) even if they could not yet provide information about the rental infrastructure (*e.g.*, number of windows). FOCUS can be extended to this case by offering users a "don't know" option for answering questions and removing unknown features from consideration in later iterations.

Another consideration for such a system is context-dependent costs—in our experiments, we assume that unit-related features are the most expensive since they may require a site visit. However, if a user is currently viewing an apartment, it becomes much less expensive to ask them to gather information for that unit since they are already on site. Finally, a real-world implementation of such a system where multiple people can potentially gather information to be shared across multiple users offers the opportunity for crowdsourcing. Unit-level information crowdsourced in this way would likely remain relevant for a fairly long time, since feature values will change only when the landlord upgrades the property. Crowdsourced information could be less expensive than in our current formulation of feature costs, but it is likely that not all units will be equally visited and reported on. Incentivizing users to provide information for under-visited rental units could be incorporated into the information-acquisition strategy.

4.4.5 Predicting student stress levels

Knowing a user's stress level at an upcoming point in time allows for automatic suggestions or reminders to help them manage stressful situations. We want to predict momentary stress reports from college students, given an assortment of data such as demographic information, depression, sleep habits, and deadlines (Wang et al., 2014). This task has redundant features with various costs. For example, we can ask a student to fill out a survey to measure their anxiety, or we can use phone sensors to measure length of sleep. The first method incurs the cost of interrupting the person (potentially at a bad time), the second the cost of draining the phone battery. Our goal is to arrive at a "good" (low-uncertainty) prediction for stress levels without exhausting too many resources by strategically selecting which features to obtain.

This application differs from the (simpler) personalized energy problem in the previous section in several key ways: (1) We want to predict *momentary* stress levels, rather than a single value constant across time, as in the energy prediction example. (2) We include a new type of cost: battery life. (3) We consider context-dependent costs: for example, when a user's phone is charging, turning on sensors is no longer prohibitively expensive.

Background

Biologically meant as a mechanism for survival, stress can become harmful if sustained for long periods of time (Moberg, 2000), with individual-level health and societal-level economic consequences (Kalia, 2002). College students face a unique and significant type of stress, partly because of their transition between dependent child and independent adult (*e.g.*, (Ross, Niebling, & Heckert, 1999)).

It is possible to estimate stress from physiological and physical signals, like skin conductance, brain activity, and pupil dilation (*e.g.*, (Sharma & Gedeon, 2012)). However, these physiological measurements often

require unwieldy, task-specific instrumentation. However, more accessible signals that could be measured in a lightweight fashion (from mobile phone data) are also associated with stress, like movement, heart rate, sleep length, and social activity (*e.g.*, (P. Ferreira, Sanches, Höök, & Jaensson, 2008; Hudd et al., 2000; Misra & McKean, 2000)).

For example, Affective Health (P. Ferreira et al., 2008) senses bodily reactions (such as movement and heart rate) and visualizes them in real time, giving users the opportunity to connect their activities to their mental state. The StudentLife project (Wang et al., 2014) collected smartphone data from 48 graduate and undergraduate students over the course of an academic term and included self-reported stress, which was correlated with other factors such as GPA[2]. Participants in the Student Life project provided nearly 1600 self-reports of stress levels, with individuals providing between 3 and 269 responses. Each self-report of stress is on a scale of 1 (least stressed) to 5 (most stressed). Figure 4.2 illustrates the number of stress reports and average stress levels for each participant.

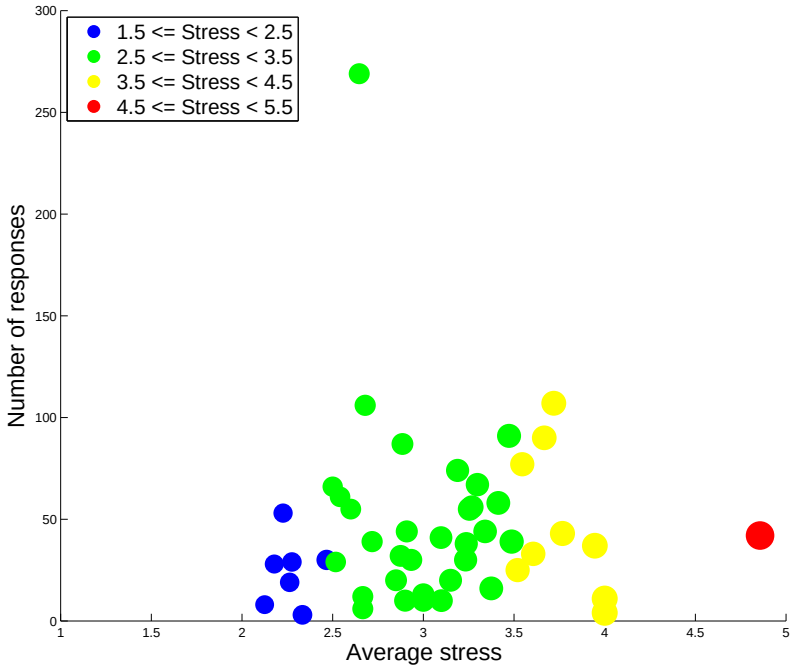


Figure 4.2: StudentLife participants' average stress (x axis and color) and number of stress reports (y axis) over the ten-week data collection period.

The StudentLife dataset is our focus as well, but we restrict our analysis to predicting stress. Our goal is to reduce the burden on users of answering experience sampling questions (the current approach to measuring stress in the StudentLife project) by substituting other features when the impact on prediction would be minimal. Thus, we use self-reports of stress at various time points as our response variable.

Cost metrics

The cost of acquiring features depends on battery drain (low for sensors like detecting light or if the phone is charging and high for sensors like the accelerometer and microphone (*e.g.*, (D. Ferreira, Kostakos, & Dey, 2015; Lu et al., 2010))) and costly interruption of the user (asking for amount of sleep or upcoming deadlines). Some of these costs are *context-dependent*: *e.g.*, if a phone is charging, battery drain from turning on mobile sensors is no longer an impediment to gathering those features. Table 4.4 summarizes

[2]The StudentLife dataset is available online at `http://studentlife.cs.dartmouth.edu/dataset.html`

the cost categories (*Low-cost sensor* measurements, more expensive *High-cost sensor* measurements, and the most expensive interruption of the *User* to provide information) of each feature and which feature costs are context-dependent.

Experimental setup

As in the energy application, we used linear regression to predict stress reports. However, rather than using feature selection to choose a subset of features from a larger set, we used previous research findings to extract features relevant to stress (*e.g.*, (Hudd et al., 2000; Misra & McKean, 2000)), along with some current context: time of day, sleep length, exercise length, length of time until next deadline, number of upcoming deadlines, current activity (stationary or in motion), current audio (silent or noisy), if the phone is currently in a dark environment, and if the phone is currently charging. We assume that time of day is freely available. We assign three additional cost categories: lower-cost sensors (*i.e.*, light detection and charging); higher-cost sensors (*i.e.*, accelerometer to measure activity and microphone to measure audio); and highest-cost user interruption (to ask questions about lengths of sleep and exercise and upcoming deadlines).

We restricted our dataset to users who provided stress, deadline, and exercise information, resulting in a total of 660 individual stress reports. We used 80% of these stress self-reports for training (*i.e.*, learning regression weights) and the remaining 20% for testing with FOCUS.

Results

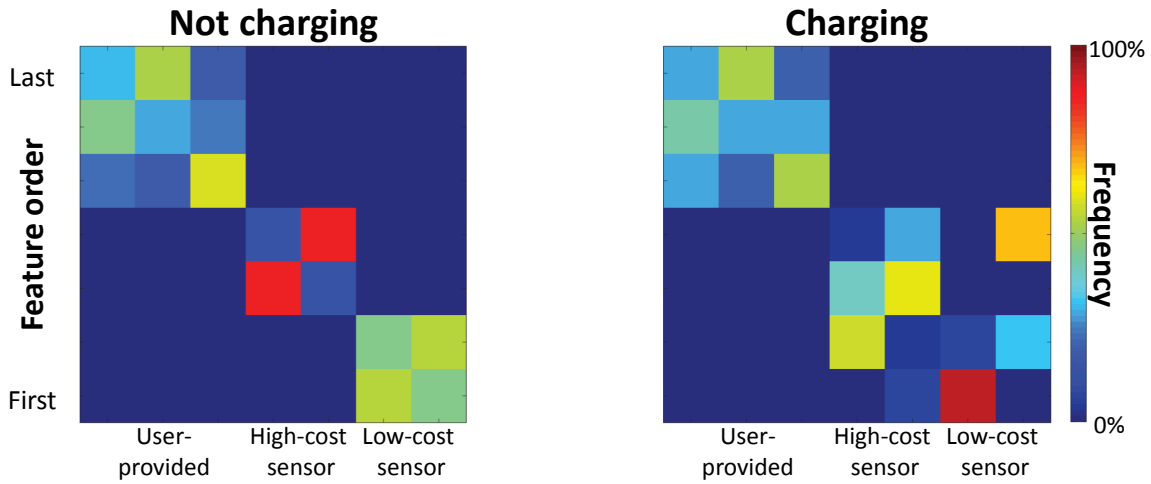
The regression model confirms several well-known properties of stress. For example, exercise is negatively correlated with stress, and number of deadlines is positively correlated with stress. Table 4.4 summarizes the predictor learned on the training set. As with the prior dataset, our approach results in predictions that are more certain (*i.e.*, have narrower prediction interval widths) than the baseline, *Fixed Selection*, while being similarly accurate. Stress prediction is shown with red lines in Figure 4.1 (which shows cost improvement of FOCUS over baseline). Accuracy levels did not differ significantly for stress predictions for any of the combinations of values shown in Figure 4.1. Prediction certainty decreased slightly for λ after $n = 5$. On average, FOCUS yielded 23% savings in cost compared to the baseline as features were added to the prediction.

Table 4.4: Information about the linear model to predict stress from sensed and user-provided data in StudentLife. The middle columns explain the cost of features (how costly, and if a feature's cost is context-dependent). The rightmost columns give the regression weights of the features in the linear model for stress prediction.

Feature	**Cost**		**Linear model**	
	Cost category	*Context-dependent?*	*Regression weight*	*p-value*
Intercept	—	—	-0.5401	0.1471
Time of day	Free	No	-0.0295	0.7950
Sleep	User	No	0.6187	0.2281
Exercise	User	No	-0.1151	0.2912
Time to deadline	User	No	-0.0401	0.8559
Number of deadlines	User	No	0.2625	0.1137
Currently moving?	High-cost sensor	Yes	-0.0226	0.8868
Currently silent?	High-cost sensor	Yes	0.0479	0.6125
Currently dark?	Low-cost sensor	Yes	-0.0864	0.2117
Currently charging?	Low-cost sensor	No	0.0492	0.5649

An important question is whether context-dependent feature costs have an impact on feature ordering. To answer this, we divided the test data into stress reports that were given when the phone was charging (20% of the reports) and those that were given when it was not plugged in. Feature order should differ in those cases, since sensor feature cost is zero when the phone is charging. Figure 4.3 shows how frequently each feature was

chosen in each position, for the charging and noncharging contexts, when cost penalty parameter λ equals 1. The battery-draining sensors are in the last four columns. When the phone is charging (Figure 4.3a), the sensed features tend to be added before the user-answered features and without regard for the relative cost of acquiring sensed features (top right 4x4 corner). When the phone is not charging (Figure 4.3b), features tend to be asked in order of increasing cost—most inexpensive sensed features first (top right 2x2 corner), followed by more expensive sensed features (middle 2x2 section), and finally by the most expensive user-provided features.



(a) When phone is not charging, sensed features are more likely to be added in order of increasing cost (the low-cost sensor features are added before the high-cost sensor features, shown by the block structure in the bottom right corner).

(b) When the phone is charging, acquiring sensor features is no longer expensive, and so sensed features (last four columns) are requested more uniformly (bottom right corner) since there is no distinction between low- and high-cost sensors.

Figure 4.3: Heatmaps showing how frequently features were chosen in each position, for the two cases for context-dependent costs: when the phone is not charging and when it is charging. Color indicates frequency; x axis indicates feature type; y axis represents feature order.

Conclusion

As in the energy prediction application, FOCUS also provided similar quality predictions for momentary stress at lower cost than the fixed order baseline. The relative cost savings of FOCUS over Fixed Selection for the stress prediction application (an average of 23%) are less than those in the energy prediction example (an average of 45%). The lesser savings for stress prediction probably occur because the stress prediction model has fewer features (9) and fewer cost categories (3) than the energy prediction model (30 features and 8 cost categories), so FOCUS has a greater variety of features with different utilities and costs to select from for the energy prediction model.

A valuable result from the stress prediction FOCUS setup is the exploration of context-dependent costs, where a feature's cost can take on different values according to the context in which the feature is being collected. Context-dependent costs are chosen differently in different settings, and FOCUS supports this flexibility that reflects the actual test-time costs of features.

4.4.6 Identifying devices in photographs

Correctly identifying a device (*e.g.*, printer, projector) in an image can support opportunistic mobile interaction with that device by automatically installing the necessary drivers without forcing the user into a manual setup. Real-time interaction speeds are crucial in this setting to make the opportunistic interaction seem truly seamless, but often image classification algorithms require computing expensive (*i.e.*, time-consuming)

features that can slow down the classification. The device identification problem can take advantage of other, less time-consuming features, like location and camera orientation to assist in prediction. User input (*e.g.*, desired use of a device, such as printing or projecting) can also inform the prediction, at the cost of user inconvenience and time.

We use a dataset of images, camera orientations, photo locations, and device capabilities (*i.e.*, "can print," "can scan", "can copy," "can fax," and "can laser-cut," for this dataset) to classify new images as particular devices (de Freitas et al., 2016). We illustrate the usefulness of FOCUS for this device classification task with decision trees and show that using FOCUS to select a dynamic subset of features to acquire at test time results in fewer expensive feature acquisitions while still correctly classifying devices.

Background

Multiple groups have considered how smartphones can be used to control physical devices, with a variety of device identification and control mechanisms. Examples include laser pointers (*e.g.*, (Beigl, 1999)), external cameras (*e.g.*, (Budde et al., 2013) with Kinect), and magnetometers (*e.g.*, (Wu, Pan, Zhang, Li, & Wu, 2010)). For the case of smartphone-taken images, past work explores directly identifying appliances labeled with fiducial markers (*e.g.*, (Liu, McEvoy, Kimber, Chiu, & Zhou, 2006)) or using image recognition to identify a pictured device (*e.g.*, (T.-H. Chang & Li, 2011)). Snap-To-It (de Freitas et al., 2016) allows users to interact with new devices by taking a picture with their phones and using both the content and context of the image to identify the pictured device and connect to it.

Snap-To-It is our focus as well, and we also use image content and context (location and camera orientation) to classify devices, as well as user input about intended device use. As in Snap-To-It, we use the Scale-Invariant Feature Transform (SIFT) algorithm to extract features from images and compare SIFT features from two images for matches—a higher number of matches means that the images are more similar (Lowe, 1999). It is possible to compute the SIFT features for the 90 reference images ahead of time, but at prediction time, the SIFT features for the image the user takes must be calculated and then compared to the reference images to check for the highest match. Our experiments show that calculating SIFT matches for a new image, against precomputed SIFT features for all 90 reference images takes, on average, 2.80 seconds, which can destroy the "real-time" feel of a service like Snap-To-It. Asking for user input is also expensive. Therefore, our goal is to give confident device predictions for images with few time-consuming SIFT match operations and overall low inconvenience on the user.

Cost metrics

We assume that image location and orientation are freely available when the picture is taken. We assign two additional cost categories for the remaining features: medium (SIFT matching) and high (asking the user about their desired use for the device (*e.g.*, "print," "laser-cut")). Although additional context could be relevant (such as whether PowerPoint is running and a projector is in the room), the Snap-To-It dataset did not include this information.

Experimental setup

The Snap-To-It dataset is pre-divided into "reference" and "testing" subsets. The reference dataset contains five images for each of 18 appliances, taken from different angles. These appliances have printing, scanning, copying, faxing, and laser-cutting capabilities. There are 108 images (six of each appliance) in the testing set. We used the reference set to construct a decision tree for image classification. Then we used this decision tree to classify test images, computing the SIFT matches and user questioning as determined to be necessary with FOCUS.

Results

We first constructed a decision tree on the Snap-To-It reference set, using MATLAB's implementation of the Classification and Regression Tree (CART) algorithm. CART is a top-down algorithm that repeatedly splits nodes of the tree (starting with all samples at the root), according to whichever binary split most decreases the "mixture" among classes in the leaves, measured by the Gini impurity (Breiman, Friedman, Stone, &

Olshen, 1984). The decision tree learned was able to optimally classify the reference set using only 7 of the 90 reference images, so we discarded the rest.

Device identification performance is shown with blue lines in Figure 4.1 (which shows cost improvement of FOCUS over baseline). On average, FOCUS yielded 29% savings in cost compared to the baseline as features were added to the prediction. In comparison to the baseline, accuracy was significantly worse only at $n = 10 - 11$ for $\lambda = .001$.

Conclusion

This application of FOCUS to real-time device identification illustrates that FOCUS improves the cost-effectiveness of feature acquisition for classification (*e.g.*, identifying a device in a photograph), as well as regression (*e.g.*, predicting energy consumption in Section 4.4.4 and stress in Section 4.4.5).

These experiments on Snap-To-It illustrate two valuable ways of cutting down time-consuming test-time image comparisons. First, using a decision tree to classify instances reduced the potential image comparison space from 90 to 7. de Freitas et al. (2016) used heuristics from image location and orientation to reduce the space of potential matches, but an algorithmic approach can expand the impact of such filtering beyond human-extractable patterns. Second, using test-time feature acquisition can further reduce the number of costly feature acquisitions on test instances that can be confidently classified without obtaining all features.

4.5 Conclusion and future work

Making real-time, personalized predictions is an important opportunity for ubiquitous computing applications. However, gathering information from users at test time can be costly, especially when not all pieces of information may be relevant for a particular user at a particular time. We have demonstrated the cost-saving value of dynamically acquiring features for test-time prediction on a variety of applications and algorithms. On all three validation datasets, FOCUS effectively lowered prediction costs (by reducing the number of additional, costly features to acquire), without sacrificing prediction quality for most values of n and λ. The FOCUS framework's ability to support context-dependent costs (illustrated in the stress prediction example on the StudentLife dataset in Section 4.4.5) allows for richer, more realistic interpretations of feature cost, which may not be fixed for all test instances.

A limitation of our work is our simplistic measure of costs for all of our predictions. A more detailed look at cost could account for real users' perception of question cost (estimated via item response times or response rates) or the exact battery drain of various sensors on the particular model of phone being used.

Furthermore, future work should explore how to connect the end-user experience to choosing a value for the cost penalty λ. This tradeoff is likely application-specific. Additionally, individual users are more or less inclined to provide information and comply with a computer system (Davis, Bagozzi, & Warshaw, 1989). Thus, determining a user-specific value for the tradeoff parameter λ would further personalize the data collection process by ensuring that the utility-cost tradeoff is appropriate for an individual. This personalization could happen by adapting the value of λ used as information is collected, based on additional information collected about the user. For example, if a user responds to information requests much more quickly than other users, that fast user may be experiencing less burden for each information request and may be more likely to comply with more burdensome information requests. Thus, a lower value of λ (*i.e.*, emphasizing item cost less in the FOCUS objective in Equation 4.2) could be used for that user.

Future work can also compare the relationship and performance of dynamic question ordering to existing paradigms of learning to make decisions, such as reinforcement learning. In Chapter 2, we pointed out that because reinforcement learning-based methods (*e.g.*, Markov decision processes (MDPs)) for test-time feature acquisition learn a policy for acquiring features at training time, it is not possible to incorporate context-dependent costs, which are not known until data collection time, into the policy. For applications without context-dependent costs, comparing DQO to an MDP could show how closely a DQO ordering matches the MDP's policy for acquiring information and the outcome trajectories (*e.g.*, cost and prediction quality as features are obtained) for the two approaches. For applications with context-dependent costs, recent work on contextual MDPs (Hallak, Di Castro, & Mannor, 2015) could be used, where the different scenarios that lead to different feature costs would be the context for the MDP.

In the stress prediction example, we considered battery charging state (*i.e.*, whether or not the phone was currently charging) as a simple binary influencer for context-dependent costs (with cost set to 0 when the phone was charging). However, more nuanced contexts could take into account the current percentage of remaining battery power, the current drain on the battery based on what applications are currently running (*e.g.*, (D. Ferreira, Ferreira, Goncalves, Kostakos, & Dey, 2013; Min et al., 2015)), or the expected time-to-next-charge (*e.g.*, (Banerjee, Rahmati, Corner, Rollins, & Zhong, 2007; Ravi, Scott, Han, & Iftode, 2008)). It would also make sense for this application to consider the influence of user context on the cost of asking them a question. For example, if a user's calendar indicates they are currently in a meeting, it may not be a good time to acquire a feature that requires user input.

Similarly, some cost metrics might take into account whether features are "shared" for multiple needs. For example, if someone answers a stress EMA multiple times in one day, the "day-level" features (*e.g.*, time spent sleeping the previous night) can be shared across predictions. Lowering the cost of such a shared feature (*e.g.*, by dividing the total cost of the feature across all instances it is expected to be used) could lead to its being selected for acquisition sooner. However, this manipulation of feature cost will need to balance the benefit lowering the cost so much that the feature is asked earlier with the risk of that costly feature's earlier appearance prompting a user to stop providing information sooner than if that feature maintained its higher cost in the FOCUS selection process.

Another aspect worth considering in test-time feature acquisition is feature confidence, especially when the same value can be obtained through different methods, with different costs and accuracies. For example, in the StudentLife case we used user-provided sleep lengths as one of the predictors for stress, but it is also possible to estimate sleep length and quality by sensing. By incorporating feature confidence into the selection criterion, we could decide whether we are confident "enough" about a value for a less costly feature to avoid acquiring a more expensive estimation of the same value.

Finally, it might be interesting to explore user- and context-specific metrics for prediction quality. Thus, the current needs of a user (in terms of accuracy) might be factored in to choices about which features are worth acquiring.

Chapter 5

Questionnaire scoring

In this chapter, we consider the problem of asking questions to measure an underlying trait, such as a personality trait, in a respondent. The final outcome of this problem (making a measurement for the latent trait) is similar to that in the previous chapter (making a prediction), but the way in which these outputs are generated differ for the two problem settings. In the prediction-making setting, a predictive model is learned that relates features (input) to a target variable (output) from training data, using methods in statistics or machine learning. Once this predictive model has been learned from training data, it can be used to make a prediction from feature values for a new test sample. In contrast, in the questionnaire-scoring setting, a formula for calculating a measurement from question responses is determined, using techniques from the social sciences (*e.g.*, (DeVellis, 2016)). Once this questionnaire-scoring formula has been determined, it can be used to make a measurement from responses for a new individual. This chapter explores question ordering for this questionnaire-scoring setting.

5.1 Introduction and background

A popular use of questionnaires is to summarize an individual's responses into a single score, for a variety of applications, such as psychology, education, and health. Often, this single score measures an underlying trait, such as extraversion, intelligence, or cultural values. The questions used to measure the underlying traits are typically repetitive, to ensure coverage of the trait. This fact can make the questionnaire to measure the latent beliefs unnecessarily long, which discourages respondents from answering all questions thoroughly. "Short form" questionnaires choose a subset of questions to ask respondents. This approach uses the same reduced set of questions for each individual, but different questions may be more informative for certain individuals than others. Dynamically choosing which question to ask next, based on previous information gathered about the respondent, can reduce the length and burden of the survey in a way that is personalized to the individual, while still gathering information about the relevant latent factors and estimating appropriate scores. In this section, we present a dynamic question-ordering scheme for assigning scores to respondents based on their responses to a questionnaire. Our approach is personalized to the individual answering the questionnaire and trades off the utility of having an answer to a potential new question with the cost of that question to select a cost-effective item.

We validate our approach for the application of assigning scores of cultural values to individuals. Our first validation is on a dataset of responses to an international survey (World Values Survey Association, 2015), in which we simulate question ordering on the complete dataset. Our second validation is a live deployment of our question-ordering method on an online experimental platform (Reinecke & Gajos, 2015). For both validation cases, we compare our dynamic question-ordering process to two baselines: a fixed-order long form that asks all questions in the same order for all individuals and a short form that calculates scores with only a subset of the full question set (all respondents receive the same subset of questions). These validations demonstrate that a dynamic question-ordering procedure can obtain estimates of equal or better quality at up to 40% burden reduction to respondents than a fixed-order long form. Additionally, the dynamic procedure allows for the possibility of asking willing respondents more questions than a fixed short form, thereby improving relative performance.

5.1.1 Summary of related work

Past work in questionnaire scoring (covered in Section 2.2) has looked at scoring questionnaire responses (*e.g.*, (Digman, 1990)), reducing survey burden by creating short form questionnaires (*e.g.*, (Stanton et al., 2002)), and personalizing tests (which can be considered a type of questionnaire) to a respondent's current score estimate (*e.g.*, (Weiss & Kingsbury, 1984)). However, there is a gap when it comes to cost-effective burden reduction for questionnaire scoring. To fill this gap, we adapt the framework of Early et al. (2016) since it iteratively trades off the *utility*, or usefulness, of the next feature to acquire against its cost. We replace their prediction-specific utility metric with a utility metric that reflects how much a new question is expected to contribute to the estimate of a questionnaire score.

Our dynamic question-ordering (DQO) approach to survey scoring is innovative compared to past work in several ways. Compared to the prevalent burden-reducing approach of short form questionnaires, DQO allows for a personalized short version for an individual. Personalized question sets often outperform a generic form applied to all respondents. DQO also offers the possibility for a respondent to continue answering questions beyond the short form, which improves the quality of the score estimate. Compared to adaptive testing approaches, DQO considers the costs of individual questions when determining which questions to ask. Questions are not always equally burdensome, and it may be more fruitful to ask a lower-cost question that may be slightly less useful than a costlier question. DQO addresses this utility-cost tradeoff. Finally, our DQO method for questionnaire scoring is lightweight in terms of data and computation requirements, in contrast to item response theory-based adaptive testing approaches (De Ayala, 2013) and Markov decision process-based approaches to feature ordering (Kakade, 2003).

Although DQO is derived from the framework presented by Early et al. (2016), this section makes several contributions over that work. First, we are considering a new problem, questionnaire scoring rather than prediction, in a new domain (calculation of cultural values). As a result, we use a new item utility metric that relates to the application of questionnaire scoring (this chapter's focus) rather than generic prediction (the focus of (Early et al., 2016)). We use empirically determined costs for items based on the behavior of questionnaire respondents rather than categorizing item costs *a priori*. Furthermore, we compare the performance of the dynamically ordered questionnaire to that of a separately developed short form, since the short form is the prevalent burden-reducing approach. Finally, we deploy a dynamically ordered questionnaire and collect data from participants, rather than only simulating the item acquisition process on existing datasets.

5.2 Survey, scales, and data

In this section, we first lay out the requirements for a survey that uses dynamic question ordering to obtain low-cost scores. The remainder of this section focuses on an example of a survey that meets these requirements, which we later use to validate our dynamic question-ordering method. We introduce the World Values Survey (WVS) and present the cultural values that this survey can measure. Finally, we develop an N-item short form questionnaire that can measure the same values using only a subset of questions from the full scale. We demonstrate that allowing each short form to be adapted to the individual (*i.e.*, each short form can select a different subset of N items for scale estimation) can achieve more accurate estimates of the cultural values than the fixed short form questionnaire that uses the same N items from all individuals.

5.2.1 Requirements

Our method for dynamic question ordering for questionnaire scoring (presented in the following section) has three main requirements of a survey.

First, it relies on historical data to learn distributions of answers. The number of historical data samples that are necessary depends on the complexity of the response distributions. In our experiments, we found that using as few as 250 training samples gave similar performance as using the full WVS dataset for training. Although this number would be high for something like a user study, data of this sort are easily available for many validated scales that are used in HCI research (*e.g.*, datasets from online personality tests[1]). Such

[1] `http://personality-testing.info/_rawdata/`

data can also be collected inexpensively using websites like Mechanical Turk, depending on how specific the study population is.

Second, the question selection criterion we use for this problem involves the contribution that a question makes to the score calculation, so the method also requires that the survey have a score that can be calculated from responses to questions. This scoring mechanism exists independently of the question-ordering procedure. Typically, domain experts develop such scales and their scoring mechanisms (*e.g.*, the User Burden Scale developed by HCI researchers (Suh, Shahriaree, Hekler, & Kientz, 2016)), and we intentionally leave scale development to these domain experts. The question selection criterion we use on the WVS dataset could apply to other surveys for score calculations, with no adjustments.

Finally, while not a requirement specific to the dynamic question ordering, respondent motivation is useful for any survey, to encourage participation.

We chose the World Values Survey (WVS) for illustration of our method due to (1) the availability of a large historical set of responses, (2) the existence of previously defined and validated scales that used WVS questions to measure individuals' cultural values, and (3) the appeal of cultural values measurement to motivate respondents to participate in our study (to receive feedback such as "Your values are most similar to those of people living in Spain and least similar to those of people living in Norway.").

5.2.2 About the World Values Survey

The World Values Survey (WVS), started in 1981 as the European Values Survey, collects people's attitudes and opinions about cultural values, across the globe (World Values Survey Association, 2015). The questionnaire is developed by an international team of social scientists and elicits respondents' opinions on politics (*e.g.*, "Would you say this is a very/fairly good or fairly/very bad way of governing this country: Having a strong leader who does not have to bother with parliament and elections?"), gender roles (*e.g.*, "Do you strongly agree or disagree: A university education is more important for a boy than a girl?"), religion (*e.g.*, "Do you strongly agree or disagree: The only acceptable religion is my religion"), technology (*e.g.*, "All things considered, would you say the world is better off, or worse off, because of science and technology?"), *etc.* Wave 6 of the WVS was conducted from 2010 to 2014 and gathered 90,350 responses from 60 countries. Figure 5.1 shows how many responses there were from each country.

The full questionnaire has around 400 questions, but some questions are country-specific and not applicable to or asked of all respondents. The data are available online in a comma-delimited ASCII file and in the formats of several statistical packages [2].

5.2.3 Measuring cultural values on the World Values Survey

Sociologists, political scientists, economists, and other social scientists have studied the relationship between economic development, democratic institutions, and cultural values (*e.g.*, (Lipset, 1959; Dahl, 1997)). One example comes from Inglehart and Welzel (Inglehart & Welzel, 2010; Welzel, 2013), who identified two types of cultural values that distinguish societies from one another. They then developed two scales to measure these cultural values using questions from the WVS. *Sacred-vs.-Secular Values* indicate to what extent a respondent departs from "sacred authority" (religion, country, group norms). The secular values scale comprises four sub-indices: defiance, agnosticism, relativism, and skepticism. Each sub-index has three component questions. For example, the "agnosticism" sub-index asks a respondent how important religion is in their life, how frequently they attend religious services, and whether or not they consider themselves a religious person. *Obedient-vs.-Emancipative Values* measure how much a person supports freedom of choice. Like the secular values scale, the emancipative values scale also has four three-item sub-indices: autonomy, equality, choice, and voice. Both measures range from zero to one, where higher values mean the person has stronger secular or emancipative values. Nations with high economic output and secular societies (*e.g.*, Sweden) tend to score higher on both scales than nations with lower economic output and religious/traditional societies (*e.g.*, Jordan). There is intra-society variation in individual scores that are related to income, education, and gender, but these contributing factors are not as predictive as nationality in characterizing an individual's cultural values.

[2] `www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`

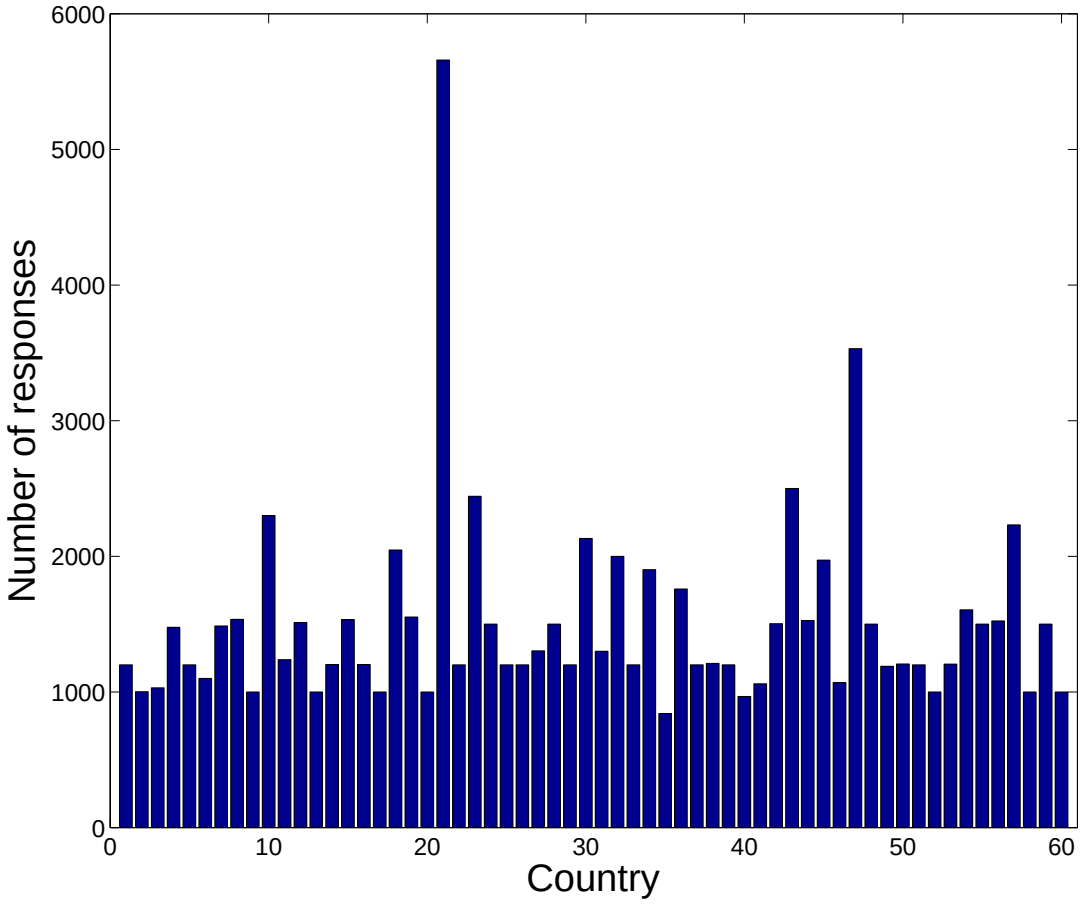


Figure 5.1: Most countries had around 1000–1500 responses to Wave 6 of the WVS. Three countries had 2500 or more responses: India (5659 responses), South Africa (3531), and Russia (2500). Three countries had fewer than 1000 responses: Trinidad (999), Poland (966), and New Zealand (841).

Inglehart and Welzel (Inglehart & Welzel, 2010; Welzel, 2013) used the WVS to identify questions and scaling formulas to measure secular values and emancipative values. Each score is calculated from responses to 12 relevant items by converting each response to a value between 0 and 1 (*e.g.*, "Strongly agree"$\rightarrow$ 0, "Agree"$\rightarrow$ 0.33, "Disagree"$\rightarrow$ 0.67, "Strongly disagree"$\rightarrow$ 1; "Yes"$\rightarrow$ 0, "No"$\rightarrow$ 1) and then averaging these converted responses. The exact process for converting question responses to scale values is detailed in the appendix of Welzel's book (Welzel, 2013).

In our experiments, we consider both the emancipative and the secular values scales from the WVS. That is, we perform all our experiments twice, once on the 12 questions that are used to calculate the emancipative values score and once on the 12 questions that are used for the secular values score.

5.3 Illustrating the benefit of adaptive forms

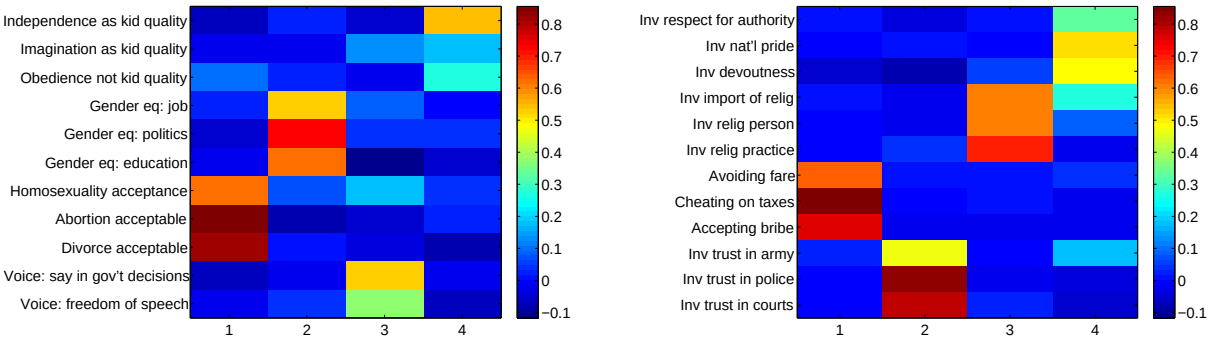
In this section, we show that allowing different respondents to answer different questions can achieve better scores than a standard subset of questions that are asked of all respondents. We first develop a short form questionnaire that asks all respondents the same subset of N questions. Short forms are the most prevalent method for estimating scores with lower respondent burden (see related work in Section 2.2.2).

Next, we compare the performance of this fixed short form to an optimal adaptive short form that also uses N questions per respondent but allows a different subset of questions for each respondent (the subset of questions that achieves the lowest-error score estimate for an individual). The optimal adaptive short form achieves better-quality scores than the fixed short form 81 – 90% of the time (and equal-quality scores the rest of the time). This result suggests that allowing individuals to answer different subsets of questions, based on their previous answers, can outperform a single short form in which all respondents receive the same subset of questions. This optimal adaptive short form requires knowing the answers to questions *a priori*, which is clearly impossible for the practical case of gathering responses from an individual. Section 5.4 will develop a practical method for choosing one question at a time, without needing to know answers to questions before those questions have been asked.

5.3.1 Developing a short form questionnaire

To identify a relevant and representative subset of questions for the short form condition, we did a factor analysis of the 12 questions in each scale of cultural values. Factor analysis aims to discover relationships among variables by expressing them as linear combinations of a latent space. That is, observed values (questionnaire responses, in our case) can be explained by the lower-dimensional space (the latent factors) (Bartholomew, Steele, Galbraith, & Moustaki, 2008). For this reason, factor analysis is often used to find latent factors that underlie a set of observations.

For the cultural values factor analysis, the models with four latent factors best reconstructed the item correlations from the full set of questions on a held-out validation set, so we used the four-factor models rather than a model with more or fewer items. Figure 5.2 is a heatmap of the factor weights for the factor analyses of the two scales (one for secular values, one for emancipative values), where the color of each cell indicates the weight of a question (y axis) on a latent factor (x axis). Warmer colors (red, orange, yellow) indicate higher factor weights, meaning that a question heavily contributes to a factor. As Figure 5.2 shows, the four-factor models make sense intuitively because questions from each sub-index (adjacent groups of questions on the y axis) score highly on separate factors. Recall that, for the cultural values scales, each sub-index consists of three related questions (such as the three questions about religion in a respondent's life for the agnosticism sub-index of the secular values scale). For the short form questionnaire, we retained the most highly weighted question in each factor (which resulted in one question per sub-index) and discarded the rest. Thus, the short form has four questions for each scale rather than the original 12.



(a) Emancipative values.

(b) Secular values.

Figure 5.2: Factor weights for all questions in each scale. It makes intuitive sense that related questions would rank highly on each factor.

5.3.2 Comparing a fixed short form to an adaptive short form

We then compared how well these four-item short forms could recreate the scores computed from the full question sets (12 questions for each scale). We also compared the performance of this *fixed short form* to

that of an *optimal adaptive short form* for each individual, to illustrate the potential benefit of adapting questions to respondents. The fixed short form is *fixed* because it uses the same four items to compute a score, for all individuals. In contrast, the optimal adaptive short form is *adaptive* because it allows any subset of four items for each individual. It is *optimal* because it achieves the lowest error for each individual. That is, it gives a lower bound on the error of a four-item scale. To find the optimal set of four items for each respondent, we calculated scores for all possible subsets of four items and chose the subset that yielded the score nearest the score obtained from the full set of items. This personalized process permits different subsets of items to be chosen for each respondent, but it requires knowing the answers for all questions.

We performed these comparisons on 10% of the WVS dataset, for both emancipative and secular values scales. Table 5.1 summarizes the performance of the fixed short form and the optimal adaptive short form. The first two rows give the mean absolute errors (with standard deviations in parentheses) for the emancipative and secular score estimates from both types of short forms. For both scales, the optimal adaptive short form yields lower-error scores, with less variance in the error among the test set, than the fixed short form. The final row of Table 5.1 presents how frequently the fixed short form (from the factor analysis) resulted in the same score estimate as the optimal adaptive short form. The fixed short form is optimal on only 9.5% and 18.9% of the test set. This frequency is higher than what would be expected by chance performance of the fixed short form (since there are $\binom{12}{4} = 495$ potential four-item subsets from which the optimal adaptive form can choose). However, these results illustrate the potential for improvement from allowing individual respondents to use a personalized short form rather than imposing a generically accurate short form on all respondents.

	Emancipative	**Secular**
Fixed short form	0.102 (0.076)	0.085 (0.070)
Optimal adaptive short form	0.009 (0.013)	0.015 (0.014)
Percentage fixed short form is optimal	9.53%	18.87%

Table 5.1: A four-item short form used for all individuals does not perform as well as an optimal adaptive four-item short form that can be personalized to the individual. The first two rows show mean absolute errors (the average of the absolute value of the difference between predicted scores and the true scores) and standard deviations for both emancipative and secular values scales. The final row shows the fraction of the time that the fixed short form achieved the same minimum error as the optimal adaptive short form.

5.4 A procedure for dynamically ordering survey questions

To take advantage of this performance improvement from personalizing the set of questions to the individual and cost reduction from using fewer questions than the full scale, we present a method that sequentially selects which question to ask a respondent next, depending on how useful that question is expected to be to the subsequent score estimate and how much that question costs. This method is an instantiation of the question-ordering framework in Chapter 3. For this application of questionnaire scoring, we want to quickly identify a subset of questions that most influence the respondent's score. Because the score is calculated as an average of values in the same interval, we can assemble the most "influential" items by maximizing the variance of the items. That is, we want to choose the next question that will maximize the variance of the answers we already have and the answer to that question. We substitute item variance as the utility measure in the DQO framework.

While different questions can influence the score in different ways, they also have different costs. At an abstract level, "cost" can be defined as the chance that a question will not be answered—for this reason, high-cost questions are risky to ask because they might not be answered and might even convince the respondent to drop out of the survey. Two easily measurable metrics for question costs are item response rates (how frequently the question is answered) and response times (how long it takes to answer the question).

In addition to considering the information from previously provided information, our question-ordering procedure will also take into account the costs of individual questions.

We adapt the test-time feature ordering framework from Early et al. (2016) to this questionnaire-scoring setting. That feature ordering procedure sequentially chooses features to acquire by trading off the *utility* of a feature, defined in terms of its usefulness to the prediction problem, with its cost. The measure of utility used for the applications in that paper was prediction certainty: how likely a prediction is to coincide with the true value. For this application of questionnaire scoring, we want to quickly identify a subset of questions that most influence the respondent's score. Because the score is calculated as an average of values in the same interval, we can assemble the most "influential" items by maximizing the variance of the items. That is, we want to choose the next question that will maximize the variance of the answers we already have and the answer to that question. For example, if a respondent has answered several items that indicate a high score value (*e.g.*, for emancipative values, they have already answered that they strongly support gender equality in politics and that they think abortion is always acceptable), it will be more "surprising" to get a low-scoring answer (*e.g.*, given their previous responses, they probably also think divorce is acceptable, which wouldn't change their emancipative score much). Thus, an answer with high variance compared to the already given answers would most change the score (since the score is calculated as a simple average of all responses). We substitute item variance as the utility measure in the FOCUS framework (Early et al., 2016).

Because this approach takes previously provided information (*i.e.*, already answered questions) into account, the utility of a candidate question q involves not only the answer for that question but also the answers to all previously answered questions. In the following notation, we will denote the answers from a d-item questionnaire in the vector $x \in \mathbb{R}^d$. We will let x_q indicate the q-th item of x (*i.e.*, the answer to question q), and $x_\mathcal{K}$ indexes the known items of x (*i.e.*, answers to all previously asked questions).

Given a set of known question responses in $x_\mathcal{K}$, computing the actual utility of question q requires knowing its answer before it has even been asked, which is not possible. Instead, we compute the expected value of the utility by taking an average of all possible outcomes (the actual utility for each possible answer $r \in R_q$ to question q), weighted by the probability of those outcomes occurring. We can calculate the utility for each potential answer r to question q by assuming the answer to q is r and calculating the utility using all previously provided answers in $x_\mathcal{K}$, along with answer r to question q:

$$\mathbb{E}[U(q)] = \sum_{r \in R_q} p(x_q = r) U(x_\mathcal{K} \cup \{r\}), \tag{5.1}$$

where $p(x_q = r)$ is the probability that the answer to question q is r and is calculated empirically from a training set (*i.e.*, $p(x_q = r)$ is the fraction of respondents in the training set that answered r to question q), and the notation $U(x_\mathcal{K} \cup \{r\})$ means calculating the utility of an answer set consisting of all the known answers (in $x_\mathcal{K}$) and the possible answer r to question q. For example, one question on the emancipative values scale is "Do you consider it especially important for children to learn independence at home?" A "yes" answer means the respondent has a higher emancipative values score than a "no" answer ("yes" is coded as 1, while "no" is coded as 0). When calculating the expected utility of this question, we would first calculate the utility of the answered questions, along with a hypothetical "yes" answer to the independence question, and weight this value by the fraction of "yes" answers in the training set. Next, we would calculate the utility of the answered questions, along with a hypothetical "no" answer and weight this value by the fraction of "no"s in the training set. The final expected value is the sum of these quantities, for the two possible answers.

The final question selection rule balances maximizing the utility of asking a new question q with minimizing the cost c_q of asking that new question, for all questions that have not yet been answered. Mathematically, the question selection criterion is

$$q^\star = \arg\min_{q \notin \mathcal{K}} (-\mathbb{E}[U(q)] + \lambda c_q), \tag{5.2}$$

where q is a question, $\mathcal{K}$ indexes known questions (*i.e.*, questions that have already been answered), $\mathbb{E}[U(q)]$ is the expected utility of q, c_q is the cost of q, and $\lambda \in \mathbb{R}$ is a tradeoff parameter. This tradeoff parameter λ controls how much importance is given to question cost in the question selection: When $\lambda = 0$, cost is not considered at all and Equation 3.1 reduces to maximizing question utility alone (this is equivalent to the

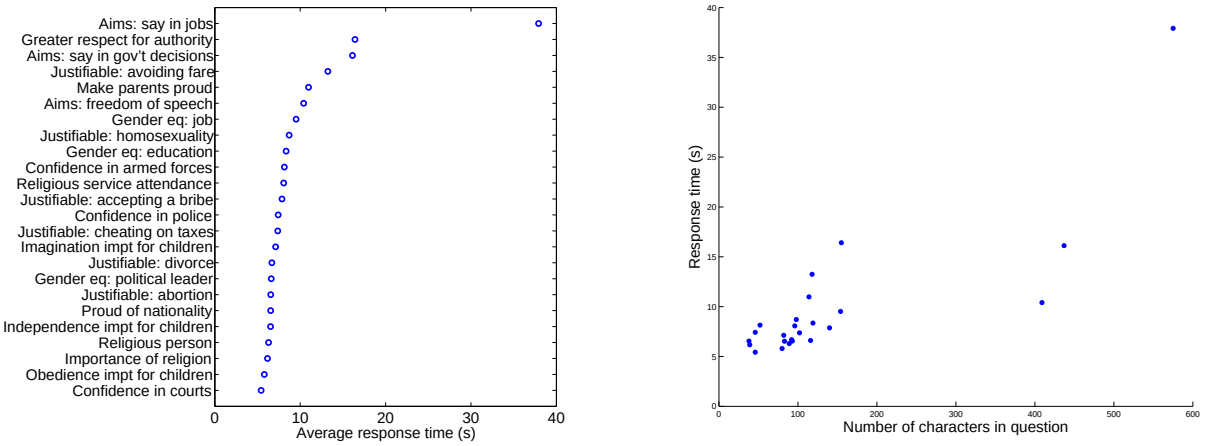
case when all questions have equal costs). As λ increases, more weight is given to question cost until the least expensive question will be chosen, regardless of its expected utility.

When we have no prior information about a respondent (*i.e.*, no questions have been answered yet), we randomly choose the initial question.

5.5 Determining question costs

Before this dynamic question-ordering algorithm can be implemented, we need to know the cost of each question. Due to the importance of completion time in respondents' decisions to answer a survey (Crawford, Couper, & Lamias, 2001), we use response time as the measure of question cost. Because the WVS dataset does not include timing information, we ran the survey on Amazon Mechanical Turk to gather response times for each question. We recruited 100 Mechanical Turk workers to take a survey with all 24 cultural values questions. Each question was asked on a separate page, and we tracked how long participants spent on each page before submitting an answer. Our estimated completion time for the survey was five minutes, and participants were compensated $1.00.

Ninety-nine respondents answered all questions. The remaining person answered 23 of the 24 questions. To calculate average response times for each question, we first removed outliers, defined to be points that were more than three standard deviations away from the mean response time for that question. We assume that these unusually time-consuming instances were due to participants' getting distracted, rather than their taking up to 25 minutes to answer a single question. Excluding these outliers, the average total time to answer all questions was 3.5 minutes (standard deviation (SD) 1.2 minutes). Average response times for each question ranged from 5.4 seconds (SD 4.0) to 37.9 seconds (SD 35.0). Figure 5.3a plots response times for the questions in the survey. These response times are correlated with the length of the question, as Figure 5.3b illustrates: Longer questions typically take more time to read and answer. Tables B.2 and B.3 in Appendix B list the costs (item response times) for the questions for the secular and emancipative values scales.



(a) Question response times.

(b) Question response times and question lengths.

Figure 5.3: Average response times in seconds for each question. Question response time is correlated with question length.

We used the average question response times as question costs for our experiments on question ordering[3].

[3]We also winsorized outliers, by setting response times above the 95th percentile for each question to the 95th percentile value. This resulted in average times *per* question that were on average only 0.46 seconds different from the outlier removal strategy and did not significantly change the results. In the remainder of the paper we report on results from excluding the outliers when calculating the mean response times.

5.6 Experiments

To illustrate the effectiveness of this dynamic question-ordering method for questionnaire scoring, we implemented it on survey responses, both historical and live, and compared the performance of DQO to a standard fixed-order long form and a standard short form. We first used a pre-collected international survey to simulate results of partial scoring under our question-ordering method. Then we deployed our questionnaire on a platform for online studies that attracts participants across the world.

To test the benefit of our dynamic question-ordering approach, we compare four conditions for collecting responses and calculating scores: (1) a *fixed-order long form* that asks all questions in a standardized order and calculates the scores from these responses, (2) a *short form* that asks a subset of questions and calculates scores from these responses, (3) a *dynamically ordered form* that asks all questions, but the order for each respondent is determined *via* our DQO method, and (4) an *adaptively ordered oracle* that asks all questions and selects the next question according to which will most reduce the error of the subsequent score (which requires knowing the answers to all questions before they have been asked). For the conditions that ask all questions, we can recover the score results at any point in the process, by calculating scores using all questions asked up until that point. The fixed-order long form was determined by taking the order in which questions were asked in the standard questionnaire form in the WVS and applying this order to all respondents. The short form gives all respondents the same four-item subsets of questions determined by the factor analysis to capture the most information in the individual responses (Section 5.3.1). The adaptively ordered oracle can choose questions in a different order for each respondent. It is an *oracle* in the sense that, since it selects the next question according to what will most reduce the actual error, it must know all of the answers to the questions before selecting which one to acquire next. The oracle provides a lower bound on score error—it is impossible for any iterative question selection scheme to achieve lower error than this oracle.

5.6.1 Performance metrics

The goal of our dynamic question-ordering approach is to quickly obtain *representative* answers that result in *accurate* scores at *low cost*. To measure these quantities, we calculated three values at each point in the question-asking process: the variance of responses provided so far, the error of the current score estimate compared to the final score, and the total cost of all questions asked so far. We get a trajectory of the values for each of these metrics at each point in the question-asking process (from when no questions have been asked to when all questions have been asked), for each test sample. We then can take the mean and variance of these metrics across all test samples to see how well the method generalizes across individuals. We can consider the metrics at each point to reflect how well a question ordering would have done, if a respondent broke off at that point in the survey.

Answer representativeness: Variance of responses

The variance of answers collected for an individual indicates how well the current set of answers captures the variation in that individual's values. Because we know the answer to a question once it has been provided, we can determine the actual variance of the answers given up to that point—*i.e.*, this metric is the actual measure of utility for the chosen question $q^\star$. The question selection criterion (Equation 3.1) uses the expected value of this quantity to choose a question to ask next, before its answer is known. Answer variance is a heuristic for the diversity of a person's score, and the question selection criterion seeks to maximize the expected variance. Thus, we would expect actual variance to be high. However, a more important metric for evaluating score quality is the error of the current estimate.

Score accuracy: Error of current score estimate

We define error as the absolute value of difference between the current estimate and the final estimate for the score. Because the final estimate is the average of the complete set of responses that have all been converted to values in the interval $[0, 1]$, we can estimate scores from partial responses as the average of responses that have been obtained up to that point. Thus, the error of an estimate y_t obtained after t questions have been

answered is

$$\text{err}_t = |y_t - y| = \left| \frac{1}{t} \sum_{q \in \mathcal{K}_t} x_q - \frac{1}{d} \sum_{q=1}^{d} x_q \right|, \tag{5.3}$$

where y is the true score, $\mathcal{K}_t$ indexes all questions that have been answered at point t, and there are d total questions. The error for a single estimate is not guaranteed to decrease as more answers are provided (consider, *e.g.*, a three-question case with answer values 0, 0.5, and 1: The final estimate is 0.5, but both question-ordering schemes starting with the 0.5 answer result in the sequence of errors 0, 0.5, 0). However, due to the often related nature of answers (since the scale is designed to measure an individual's values along a single dimension), errors tend to decrease as questions are answered.

Cost of questions asked

The successive cost is defined as the sum of the costs of all questions asked up to the current time. Because all costs (item response times) are positive, this metric increases as more questions are answered.

5.6.2 Questionnaire scoring and question ordering on the World Values Survey dataset

We began by simulating the question-asking and -answering process on the pre-collected World Values Survey dataset by hiding the answers to questions until they were selected in the question-ordering process and "asked." Once a question was "answered," we unveiled its value. We used 90% of the 90,350 responses as training data, to calculate empirical probabilities of question responses and to determine the fixed-order baseline. On the remaining 10% of the dataset, we considered responses individually and applied our question-ordering procedure to choose questions to ask.

Figure 5.4 plots the error, cost, and variance as questions are answered for emancipative values (left column) and secular values (right column), for our dynamically ordered questionnaire (DQO) with a variety of cost penalties λ, the oracle, and the two baselines: the fixed-order long form and the four-item short form. For the methods that use all questions (DQO, Oracle, and Fixed order), the lines meet once all 12 questions have been answered because then all the methods have the same information.

The first row of Figure 5.4 shows that the dynamically ordered questionnaire tends to have lower error than the fixed-order long form, particularly when few questions have been answered. The dynamically ordered form reaches similar error as the four-question short form once four to six questions have been answered. Beyond that number, the long forms usually have lower error than the short form since they are gathering more information. The oracle almost always has the lowest error, due to its omniscient error-minimizing selection rule. The oracle's errors near the end of data collection sometimes exceed those of the other methods due to the greedy, single-question lookahead, question acquisition rule of the oracle (see Section 5.6.1 for an example).

The middle row of Figure 5.4 shows that, as expected, costs are lower when higher penalties λ are placed on question cost in the DQO item selection rule. Particularly when cost is penalized, the dynamically ordered questionnaire has lower cost than the short form, even when the dynamic form asked more than four questions (the number in the short form). In the emancipative values calculation, the oracle has a similar cost trajectory to the methods that do not strongly penalize question cost. In contrast, the oracle for the secular values calculation has higher cost than the other question orderings. This illustrates that error-reducing questions are sometimes more costly (as happens for the secular values), but not always so (emancipative values).

The last row of Figure 5.4 illustrates the variance of items as questions are asked. Due to the selection criterion of DQO, the items most expected to increase variance are chosen near the beginning, with overall item variance decreasing as more questions are answered. Variance is highest when there is no or little penalty on question cost, because then the variance-maximizing term dominates the question selection rule in Equation 3.1. It makes sense that the oracle would have low variance, since the oracle sequentially selects questions that result in low-error score estimates—the oracle first chooses the single question whose numeric value is nearest the overall score (the mean of all answers) and continues to add questions that yield a mean near the final score.

These results on the simulated dataset illustrate the usefulness of DQO, in terms of score error and cost to the respondent: DQO yields lower-error score estimates with few questions than the fixed-order long form, while maintaining a lower cost, especially for higher cost penalties λ. Additionally, DQO allows for a willing respondent to continue answering questions beyond the number in the short form, which lowers the score estimate error compared to the short form. It remains to be seen how respondents will respond to a survey with dynamically ordered questions, which is the setting our next experiment addresses.

5.7 Questionnaire scoring and question ordering: Live on LabintheWild

To assess the effect of question ordering on survey respondents, we deployed a live implementation of a survey with our dynamic question-ordering algorithm to a population of Internet users. Since the simulation results showed that $\lambda = 0.01$ achieved lower-cost estimates with no decrease in performance, we used this value for the cost tradeoff parameter in the deployment. Our deployment platform was LabintheWild (Reinecke & Gajos, 2015), which recruits Internet users from across the world to complete studies. LabintheWild participants are not financially compensated but instead receive personalized feedback based on their responses.

Our study attracted 853 participants from 67 unique countries. 48% of respondents were from the United States. Table 5.2 summarizes participants' countries. 50.2% of these participants were female, and the average age was 28.4 (SD 11.8).

Country	**Number of responses**	**Percentage**
United States	431	48.3%
United Kingdom	53	5.9%
Canada	48	5.4%
Australia	45	5.0%
Finland	23	2.6%
Sweden	20	2.2%
Germany	19	2.1%
Netherlands	19	2.1%
Singapore	16	1.8%
France	12	1.3%
Spain	12	1.3%
Others	194	21.7%

Table 5.2: Participant countries for the LabintheWild deployment.

5.7.1 Implementing a dynamically ordered questionnaire

We implemented our dynamic question-ordering algorithm on the SurveyGizmo platform[4]. SurveyGizmo's "professional" subscription option offers a "custom scripting" feature, using php-like SurveyGizmo functions. These functions support retrieving answers to questions, determining which questions have been answered, and jumping to a specified page of a survey, among other functionality. Our implementation in SurveyGizmo involved creating a script to run between pages. This script (1) pulls the responses to previously answered questions, (2) calculates the expected value of each question that might be asked next (from Equation 5.1), (3) combines the expected value of each feature with its cost and selects the optimal question to ask next (Equation 3.1), and (4) jumps to the page of the next question. Each question was displayed on its own page for two reasons: (1) to allow the jump-to-page SurveyGizmo function to achieve the question ordering and (2) to allow us to calculate response times for individual questions. The full survey contains 24 questions used for determining cultural values (12 questions for each of the secular and emancipative values scales) and an additional 15 demographic questions.

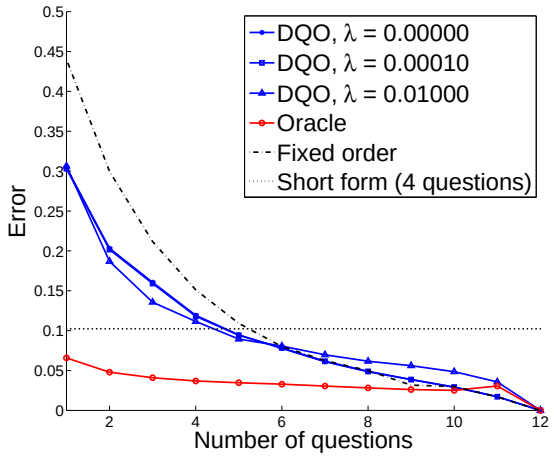
[4] `http://www.surveygizmo.com`

On the final page of the questionnaire, we gave the respondent feedback on their cultural values scores, in terms of what country they were most and least similar to, overall and for the secular and emancipative values separately. We explained the meaning of the two scales and told respondents if they scored low, in the middle, or high on each scale and gave an example answer they gave that contributed to their score on that scale. Respondents also had an opportunity to give us feedback about their survey-taking experience or thoughts on their results in a comment box on the last page.

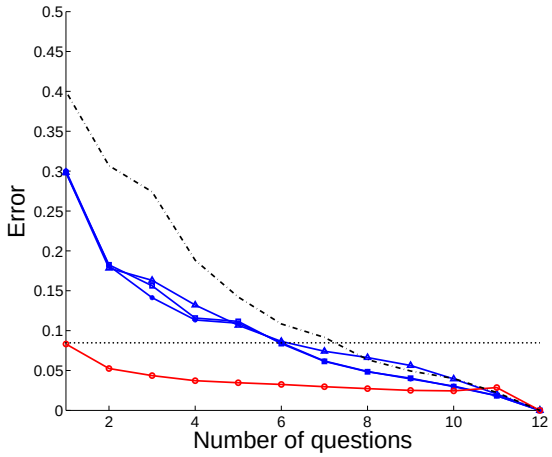
5.7.2 Results of deploying the survey to LabintheWild

With the live deployment, we can analyze the performance trajectories as we did for the WVS (Section 5.6.2) and also the feedback that participants gave after completing the questionnaire and receiving their cultural values scores. Figure 5.5 plots the error, cost, and variance metrics for the LabintheWild results, for the question orders given by DQO (with $\lambda = 0.01$ for the deployment) and the two baselines: the short form and fixed-order long form. As in the previous section's WVS simulation, DQO yields score estimates that cost less than the fixed-order long forms while maintaining similar accuracy. Furthermore, DQO can achieve lower error than the short form, if a respondent is willing to answer a few more questions than are in the short form. While these overall trends hold across both WVS and LabintheWild validations, there are some differences. In the WVS dataset, the emancipative values error trajectory (Figure 5.4a) was higher for the fixed-order long form (dashed black line) than DQO (solid blue lines). In the LabintheWild dataset, the fixed-order long form and DQO yield similar emancipative values error trajectories. This behavior may be due to the difference in score distribution among responses in the WVS dataset (emancipative values: 0.41 ± 0.18, secular values: 0.36 ± 0.17) and responses in our LabintheWild dataset (emancipative values: 0.72 ± 0.15, secular values: 0.61 ± 0.17). These differences are caused by the overrepresentation of the United States in our deployment.

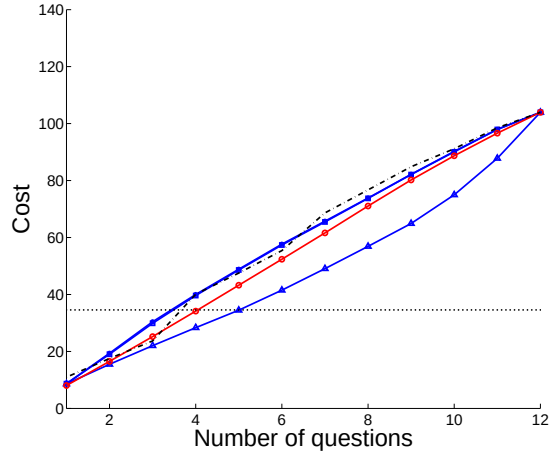
On the final page of the survey, on which respondents received their results and explanations, there was an optional free-response question where they could give feedback on their survey-taking experience. Most respondents left this field blank. The 33 responses for feedback mostly fell into the following categories: 8 respondents commented that the survey was "interesting" or "great" (*e.g.*, "Very interesting! I like that you shared connections between viewpoints."); 4 requested more information about scores (*e.g.*, "I want to know how my score compares to the country I live in...Also, maybe a little bit more about Slovenia, which I had closest values for in both cases? I mean, there's always google but right now it's just a word and it takes a little research to get the 'taste' of a country, you know?"); 5 clarified their answers (*e.g.*, "I'm a new international student who just arrived a new county for about 2 months. In the case, the question asked which country I am currently living in is kinda biased, as I am not sure what to actually fill in."); 5 suggested new questions (*e.g.*, "I think there should be more questions about: 1. Family, how important are your family and friends and how much do you trust family over friends. 2. Food, how important is food for you, is there a better cook then your mother ? Do you rather eat out or stay in. 3. Social, do you hang out with mainly boys, girls or a mix. Do you prefer to stay at home with friends or got out. 4. Man vs woman; who do you think is more intelligent / more likely to give up / who do you trust more as for instance your driving instructor?"); 3 reflected on how the questions could be interpreted and answered differently from different people (*e.g.*, "It seems that some of the questions are not 'pure' measures of cultural values—for example, the faith in police force question would depend heavily on the actual PD where the person lives. Two people with identical cultural values might answer very differently if one of them lives in a city with a corrupt police department."). Figure 5.6 shows the most common words in the feedback.



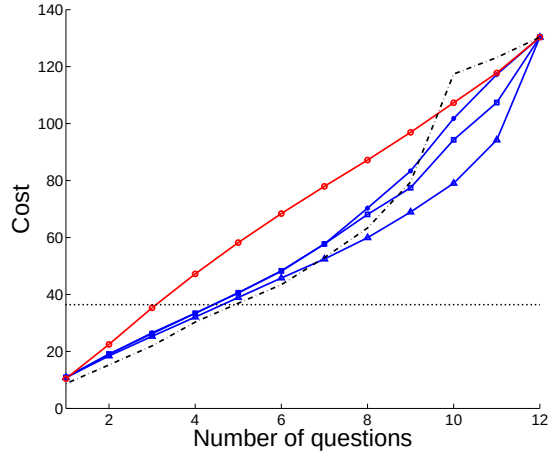
(a) Emancipative values: Successive error.



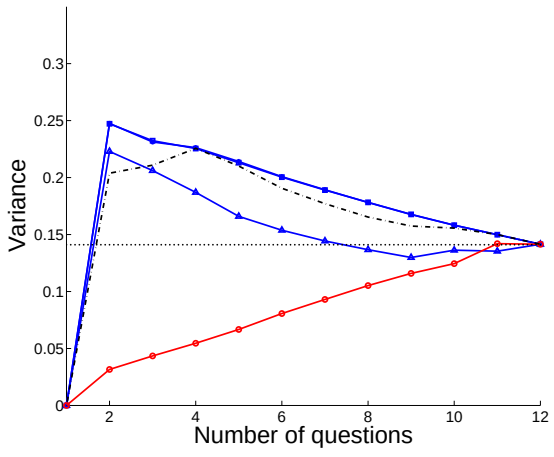
(b) Secular values: Successive error.



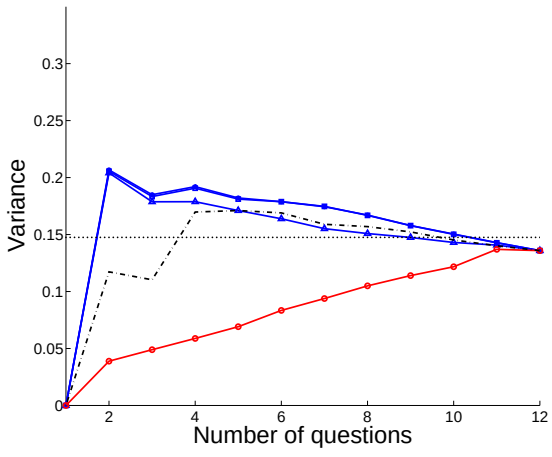
(c) Emancipative values: Successive cost.



(d) Secular values: Successive cost.

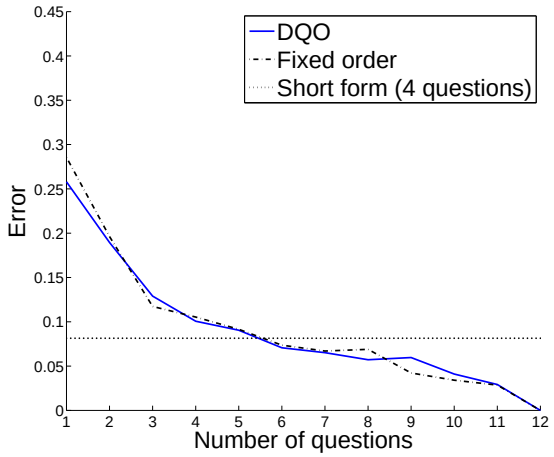


(e) Emancipative values: Successive variance.

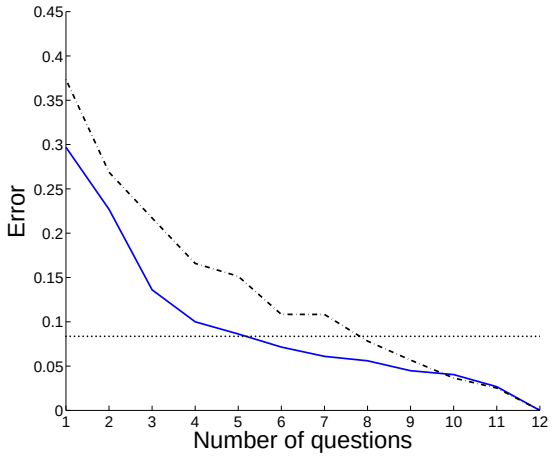


(f) Secular values: Successive variance.

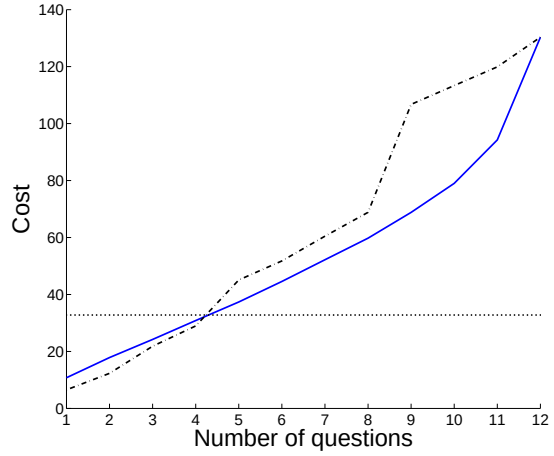
Figure 5.4: Results from simulation with the WVS dataset: Plots of successive error, cost, and variance as questions are asked, for our dynamic question-ordering procedure (DQO, with several cost penalties λ) as well as two baselines, a fixed-order long form and a short form that asks only four questions.



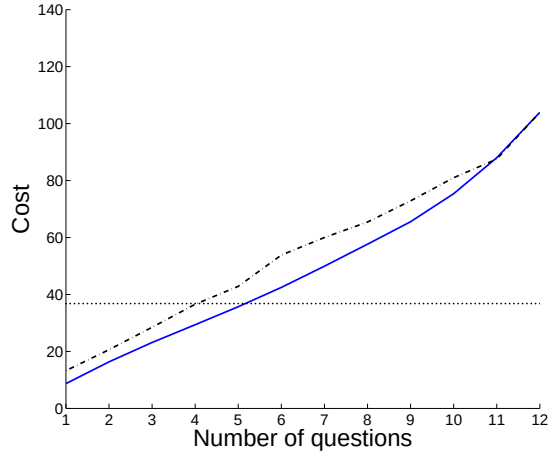
(a) Emancipative values: Successive error.



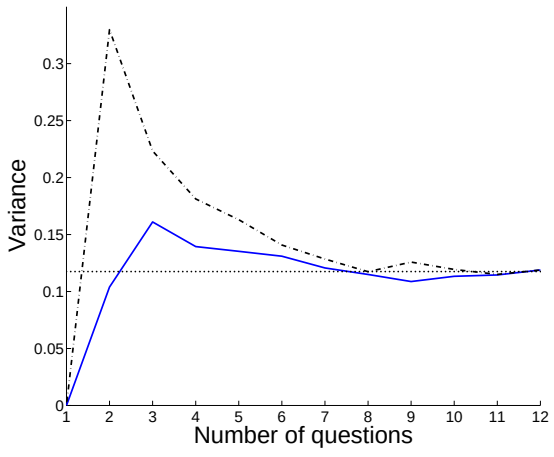
(b) Secular values: Successive error.



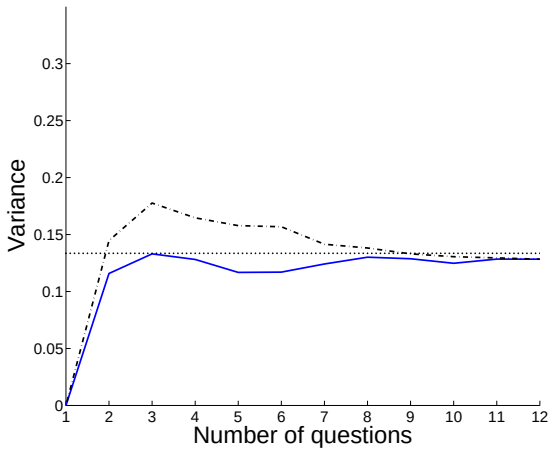
(c) Emancipative values: Successive cost.



(d) Secular values: Successive cost.



(e) Emancipative values: Successive variance.



(f) Secular values: Successive variance.

Figure 5.5: Results from live deployment on LabintheWild: Plots of successive error, cost, and variance as questions are asked, for our dynamic question-ordering procedure (DQO) as well as two baselines, a fixed-order long form and a short form that asks only four questions.

Figure 5.6: A word cloud of the most common words participants used in their responses to an optional free-text field for survey feedback.

Respondents gave no indication that they found the dynamically ordered form confusing or disorienting. While the questionnaire did have related and similar questions that would have been presenting sequentially in a fixed-order form (*e.g.*, asking all religion-related questions in a row), it is possible that these questions had low enough cognitive burden (Tourangeau, 1984) that it did not trouble participants to skip among unrelated questions. This aspect of the feasibility of dynamic question ordering is likely application-dependent—some questionnaires have sections of questions that are very related and easier to answer if such questions are grouped together.

5.8 Conclusion and future work

Being able to quickly understand users' questionnaire responses, with low burden on users, is important for researchers who want to understand participants' attitudes and values. In this chapter, we demonstrated that, unlike previous approaches that try to decrease cost uniformly across all participants (by asking everyone the same subset of questions in a short form), allowing the subset of questions to be adapted to the individual can improve performance. We developed a method that leverages previously provided information from a user to select which question is best to ask next, trading off the expected usefulness of that question against its cost. This method requires a predetermined mechanism to calculate scores from responses and a relatively small set of training data that captures distributions of responses.

Our experimental results illustrate the benefits of DQO in several ways. First, allowing each respondent to receive a personalized set of questions yields improvement over using the same subset of questions for all respondents, as is the approach of using short forms. Our experiments comparing the score performance of standard N-item short forms to personalized N-item short forms found that the personalized form outperforms the uniform short form for up to 90% of respondents (Section 5.3). Because this comparison required considering all possible subsets of N items to determine the "optimal" adaptive short form, we developed a feasible heuristic for iteratively selecting questions to ask. With this iterative approach, it is not necessary to know the entire budget (*e.g.*, number of questions, amount of burden) upfront. Our question selection rule (Equation 3.1) is based on the intuition that the most useful question to ask next is the one

that is likely to change the subsequent score. This question selection rule also incorporates a penalty on question cost, since it may be more beneficial to ask a question that has lower cost but is (slightly) less useful than a costlier question (especially if that costlier question might prompt the respondent to break off from the questionnaire). In the simulated experiments on 9,035 responses from the World Values Survey dataset (Section 5.6.2), we found that DQO has lower error (on average, DQO outperforms fixed-order long forms until 50-70% of questions have been asked) and lower cost (up to a 40% cost reduction, depending on cost parameter λ) than the fixed-order long forms. A live deployment that gathered 853 responses on the LabintheWild platform yielded similar results for DQO outperforming the baseline of the fixed-order long form (Section 5.7.2). Furthermore, the dynamic procedure allows the questionnaire administrator to continue asking questions beyond a short form's capacity, when a respondent is willing to spend more time or effort answering questions, resulting in improved score estimates. These results are likely to be even more pronounced in longer surveys, where breakoff is more likely (Peytchev, 2009).

The DQO approach presented here can be widely applied to any questionnaire that results in a score for each respondent (cultural values, in this application). The method requires historical data to learn distributions of question responses. We found that as few as 250 training samples achieved similar performance as using 90% of the WVS dataset (approximately 80,000 responses) to learn these distributions. In the case where precollected training data are not available, it would be relatively inexpensive to collect a sufficient number of responses using a crowdworking platform like Mechanical Turk. DQO is ideal for the questionnaire-scoring case where the accuracy of the short form is too low and the cost of the long form is too high. If respondents answer all questions anyway, the order in which those questions are asked does not matter (except for the case where the order of questions affects how respondents interpret and answer questions (McFarland, 1981)).

In this application, we validated our dynamic question-ordering approach on a survey with two scales to calculate two types of cultural values of the respondent: secular and emancipative values. We treated these two scales individually, first dynamically ordering the questions associated with the first scale and then dynamically ordering the questions associated with the second scale. However, a more nuanced approach for multiple-scale questionnaires could choose items from the different scales in tandem, rather than doing one scale at a time. Often, scales in the same survey are related, and we can take advantage of these relationships to achieve even better cost performance. We could incorporate multiple scales into the item selection rule by learning relationships between scores and all questions (not just the questions that explicitly define each score) and modifying the utility function so that each question's utility includes its effect on *all* scales. Incorporating this additional information will further reduce the burden on respondents (since all questions can influence all scores), particularly for longer dynamically ordered questionnaires compared to the longer fixed orders.

A topic for additional personalization is considering respondent-specific values for the cost tradeoff parameter λ. Questionnaire burden is a subjective phenomenon (*e.g.*, (Bradburn, 1978; Yu et al., 2015)), and some individuals are more willing than others to spend time and effort answering questions. Adapting the value of λ to an individual, based on their behavior in answering the first few questions or background information (*e.g.*, demographics) that is prior knowledge, can ensure that the utility-cost tradeoff makes sense for the individual who is providing information.

Another extension is to use prior information to initialize the first question to ask, rather than randomly selecting it (as we did). For example, an online survey could identify a respondent's location (country or region) from their IP address and initially select a question that is most likely to differentiate respondents in that country. Furthermore, if the questionnaire is part of a series (since many online research sites offer multiple questionnaires, *e.g.*, (Reinecke & Gajos, 2015)), prior information from a respondent's previous surveys can inform the choice of first question.

Additionally, we can consider information that can be automatically collected while the user is taking the survey, along with the information the user provides. For example, *paradata* collected as the respondent answers an online survey (*e.g.*, time spent on page, mouse clicks (Kaczmirek, 2008)) can be used to model user engagement, with survey action taken to increase engagement and response rates (M. P. Couper et al., 2010).

5.9 Summary: Question ordering for single-output data collection

This chapter and the previous chapter explored ways to perform adaptive data collection when the purpose of collecting data is to make a single assessment from the collected data. First, in Chapter 4, we considered prediction in the classical machine learning sense, where adaptive data collection is performed when gathering feature values, to make a prediction on a new test point. Then, in this chapter, we considered questionnaire scoring, where adaptive data collection is performed when asking a respondent questions, to measure a latent trait. In both cases, we considered the mechanism for prediction-making to be fixed before the adaptive collection procedure began—*i.e.*, the predictive model and questionnaire-scoring rules had already been determined from preexisting, complete training data. We then sequentially used the impact a new feature or question would have on the final assessment, trading off this utility against the cost of each potential piece of information, to determine which piece of information to acquire next. In the following chapter we consider how to perform adaptive data collection when there is no single prediction or measurement to be made as the goal of the data collection.

Chapter 6

Data collection-focused question ordering

In the previous two chapters, the goal was to gather information for making predictions and we saw that not all information was necessary to make a good-quality prediction. Therefore, dynamically ordering features to acquire could move the most useful (in terms of high utility and low cost) to the beginning of data collection and not expend the resources to acquire all features. In contrast, surveys are typically concerned with gathering *complete* information from a population. However, survey respondents are not always willing to take the time or effort to fill out complete surveys, as evidenced by declining response rates (*e.g.*, (Porter, 2004; Shih & Fan, 2008)) and breakoffs partway through surveys (*e.g.*, (Horwitz, Tancreto, Zelenak, & Davis, 2012)). Thus, there are two scenarios for survey-focused dynamic question ordering that we consider: (1) modeling user engagement and ordering questions in ways that will keep users motivated to complete the survey, and (2) collecting the information that most characterizes the respondent so that if they do drop out of the survey, imputed values for their unanswered questions will be accurate. Furthermore, some surveys do have prediction goals (*e.g.*, predicting if a respondent is in the labor force, for the employment rate estimate in the Current Population Survey (U.S. Bureau of the Census, 2006)), so the prediction-guided question ordering from the previous chapter could also apply to these scenarios.

Any statistics-based approach to dynamic question ordering of the sort we consider here would seem to run counter to traditional arguments that questionnaires should have a fixed structure for all respondents and when the same quantities, *e.g.*, unemployment or poverty, are measured by surveys over time. Just over thirty years ago, the cognitive aspects of survey methodology (CASM) movement, *e.g.*, see (Jabine, Straf, Tanur, & Tourangeau, 1984; Sudman et al., 1996; Tanur, 1992), made the argument that this traditional approach to survey design shackled respondents and often prevented them from providing the very answers that the survey methodologists sought for their questions, *e.g.*, see (Suchman & Jordan, 1990; Tanur, 1992). We believe our approach reopens the door to the arguments raised by that movement, but in a very different manner, and somehow survey statisticians will ultimately need to blend the lessons from the CASM movement with the needs for cost-driven dynamic ordering. One important problem is that order effects (Sudman et al., 1996) can influence how people interpret and answer certain questions, leading to inconsistent data across respondents who received questions in different orders. However, not all questions are susceptible to order effects, and it is possible to constrain dynamically ordered questionnaires to respect relative orderings among sensitive questions.

In this chapter, we consider two large-scale surveys conducted by the U.S. Census Bureau: the American Community Survey (ACS) and the Survey of Income and Program Participation (SIPP). Because such surveys have multiple goals and complex designs, we begin by looking at each aspect of dynamic question ordering separately: ordering questions to predict a survey outcome, ordering questions to improve imputation quality of unanswered questions, and ordering questions to improve user engagement and response rates.

6.1 Moving from population-level adaptation to respondent-level adaptation

Typical adaptive survey designs are adaptive at the population level—they choose what survey protocol to use between phases of data collection (*e.g.*, (Groves & Heeringa, 2006)); they decide how much effort to expend on making contact for a survey (*e.g.*, (Beaumont et al., 2014; Calinescu, Bhulai, & Schouten, 2013)), sometimes using paradata about contact attempts (*e.g.*, (M. Couper & Wagner, 2012; Lundquist & Särndal, 2013; Sauermann & Roach, 2013)); they assign subsamples to fixed subsets of questions (*e.g.*, (Gonzalez & Eltinge, 2008)). Dynamically ordering questions *for a single respondent*, based on their previously answered questions, can also be considered an adaptive survey design. It is this aspect of respondent-level adaptation that we are studying.

As in the previous chapter on predicted-guided question ordering, we consider iteratively asking questions to a single respondent, trading off how useful the question is against how costly its answer is to obtain. Whereas in Chapter 4, a feature's utility reflected how useful it was to the subsequent prediction, here we define a question's utility in terms of how it influences the respondent's engagement or the quality of imputations for unknown values.

6.1.1 Using paradata to model user engagement

Information collected about the survey process, known as paradata (originally "process data" (M. P. Couper, 1998)), can indicate a respondent's engagement with and understanding of the questionnaire they are answering. Paradata can include interviewer-collected data like response times, changing answers, returning to earlier questions in the survey, clicking help buttons, *etc.* Past work in paradata analysis has focused on how paradata reveal respondents' survey-taking process, how paradata can identify usability issues in surveys, and how paradata can inform population-level adaptive design. For example, Bassili and Fletcher (1991) andHeerwegh (2003) showed that participants who take longer to answer questions about attitudes also tend to change their minds when presented with counterarguments to their original response. They conclude that people with less attitude stability need more time to come up with an answer than people who are already confident in their attitudes. Paradata about users' interactions can also reveal that a survey instrument is not designed well, as Healey (2007) illustrated by finding that drop-down menus in online surveys increased item response times; question length (*e.g.*, (Bassili & Scott, 1996; Yan & Tourangeau, 2008)) and complexity (*e.g.*, (Wagner-Menghin, 2002; Yan & Tourangeau, 2008)) also increase response times. Paradata have been used for adaptive survey design too; in particular, for predicting the likelihood that a unit will be interviewed on a next contact attempt (Groves & Heeringa, 2006; M. Couper & Wagner, 2012) and for analyzing differences in population estimates as contacts are made (Lundquist & Särndal, 2013).

6.2 Application to the Survey of Income and Program Participation

Our first application, on the Survey of Income and Program Participation (SIPP), considers ordering questions to predict a survey outcome and ordering questions to improve imputation quality of unanswered questions. Conducted by the U.S. Census Bureau, SIPP collects data on income, employment, and social program participation and eligibility from households (U.S. Bureau of the Census, 2014b). The SIPP is designed as a longitudinal national panel survey, where each panel is a representative sample of 14,000 to 52,000 households, contacted yearly for three to five consecutive years. Each household interview is conducted in person, via a computer-assisted personal interviewing (CAPI) instrument, and aims to get self-reports from all household members at least 15 years old. In addition to demographic information, interviews ask respondents for their participation in various social programs, financial situation, and employment status, in the previous calendar year. The chief goal of SIPP is to understand household program eligibility and participation and to assess the effectiveness of social programs like Supplemental Security Income, Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Families, and Medicaid. Using participation in each program of interest as the prediction of interest could guide a DQO approach.

The unique design of SIPP allows us to explore four facets of our DQO framework. First, we determine individual question orderings to optimize response quality (survey completion and imputation quality). Second, we determine question orderings to optimize the quality of predictions of respondents' participation in each social program of interest. It is very likely that not all questions are necessary to predict participation, and DQO could reduce respondent burden (interview length) while still collecting the information the Census Bureau needs. Third, we consider the effect of this respondent-level adaptation on population-level estimates. Finally, we use the longitudinal nature of SIPP: data from prior interviews can be used as "background" information in subsequent interviews.

6.2.1 Data

In our experiments, we use the SIPP Synthetic Beta (SSB), a synthetic dataset created by first multiply-imputing missing values in the SIPP Gold Standard File and then multiply-imputing replacement values for actual responses to preserve privacy (Benedetto, Stinson, & Abowd, 2013). The SSB combines SIPP data with administrative records from sources like the Internal Revenue Service. These data from administrative records are not asked of a respondent and can be considered initializing information when selecting questions to ask. A downside of the SSB is that household structure is not preserved—each respondent is treated as a separate entity. Though there are spouse links for female-male married couples, this information is insufficient to recover a household structure which may include same-sex married couples, unmarried partners, children, other relatives, and other household members.

Defining question costs

In this case study, we use item nonresponse rates as a proxy for question burden. We calculate item nonresponse rates as the fraction of respondents for whom a value to a particular survey item had to be imputed, using publicly available SIPP data[1]. We consider data from administrative records in the SSB to be "free," since no household had to answer questions to get those values.

6.2.2 Prediction-guided question ordering

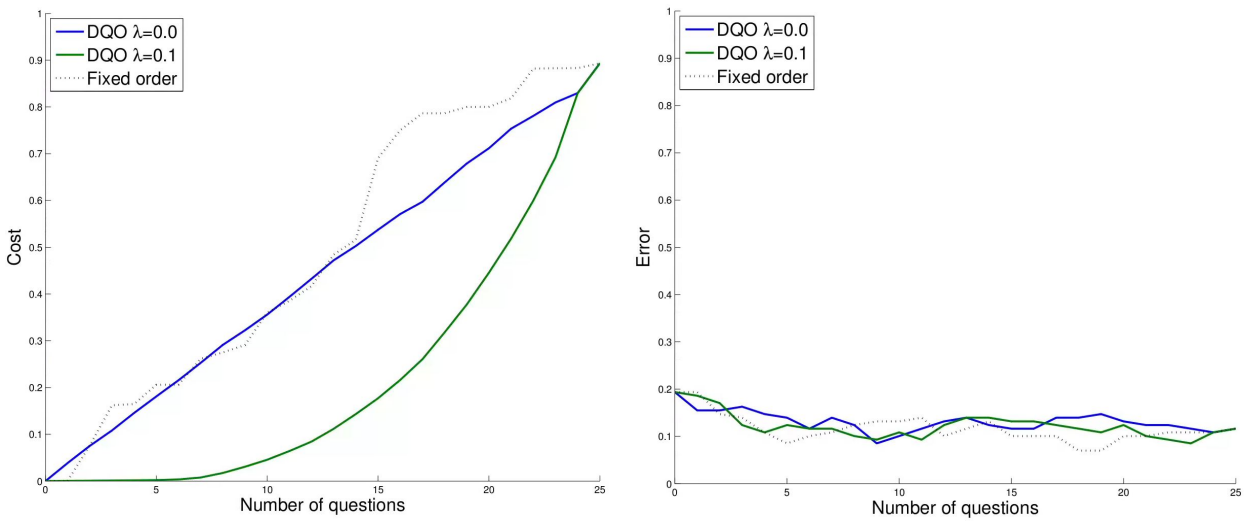
Here we consider the scenario in which a user is providing information to receive a personalized prediction based on the information they provide. We assume a predictive model has already been developed from a training set, and we want to make a prediction on a new test point. In this setting, it is likely that not all features will be needed to make a reasonable prediction. Therefore, we want to ask the most cost-effective set of questions that will maximize prediction quality while minimizing collection costs. When features are being collected to refine a prediction, a natural utility metric to use for test-time feature selection (Equation 3.1) is one that reflects how an additional feature f contributes to the *quality* of the subsequent prediction. Here we define the utility of a feature to be the *certainty* of the subsequent prediction made using all known features plus the newly acquired feature.

For this setting in the SSB, we took a respondent's participation in a program (food stamps) as the prediction of interest to guide the selection of questions for a particular individual. We compared trajectories of prediction cost (sum of the costs of all questions asked up to a point) and prediction error (fraction of test samples that were incorrectly predicted, based on the information acquired up to that point) for DQO and a fixed-order baseline that acquired features in the order in which questions are asked in the SIPP interview. Figure 6.1 shows that DQO yields lower-cost predictions at similar accuracy to the fixed-order baseline.

6.2.3 Entropy-based question ordering

A prediction-guided approach to dynamic question ordering, as in Section 4, is only reasonable when the goal of information gathering is to make a prediction. Now we consider how to dynamically order questions in a way that most characterizes the respondent so that if they do drop out of the survey, imputed values for their unanswered questions will be accurate. To measure the amount of information expected to be contained in answers to potential questions that might be asked next, we use *conditional entropy* as the measure of

[1] `http://thedataweb.rm.census.gov/ftp/sipp_ftp.html`



(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed-order baseline.

(b) DQO has similar prediction error as the fixed-order baseline.

Figure 6.1: Charts showing impact on cost and prediction error (y axis) of λ and number of features, for DQO (solid lines) and the fixed-order baseline (dashed lines).

utility. Entropy measures the information in a random variable X. If this random variable can take on one of D possible values, its entropy is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{6.1}$$

Conditional entropy measures the information in a random variable Y, when the value of another variable X is already known:

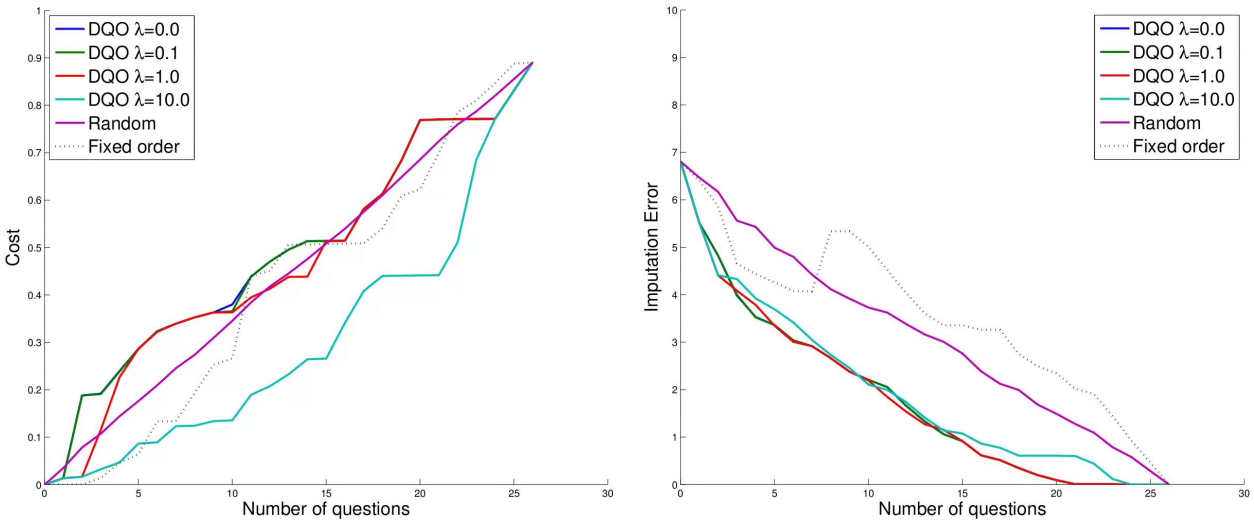
$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x). \tag{6.2}$$

In the experiments in this section, we used conditional entropy to measure question utility for the DQO question selection rule (Equation 3.1). We ordered questions and calculated performance metrics for the DQO-ordered questions with a variety of cost penalties λ, a random order selected for each respondent, and a fixed-order baseline according to the order in which these questions are asked in the SIPP interview. At each point in the question-asking process, we calculated the cost and imputation error. Cost is defined as the sum of the costs (nonresponse rates) for all questions asked thus far. Imputation error is defined as the number of incorrect imputed values for yet-unanswered questions. We imputed values for unanswered questions by using k-nearest neighbors (k-NN) to find the points in the training dataset nearest to the current respondent and predicting unknown values as the mode of those values among the nearest neighbors. This is not the method used for imputation in the SIPP (U.S. Bureau of the Census, 2014b).

Figures 6.2 and 6.3 show the costs (Fig. 6.2a and 6.3a) and imputation errors (Fig. 6.2b and 6.3b) as questions are asked in the SIPP for the DQO-ordered sets, a random question ordering, and the fixed-order baseline.

Long-term survey collection

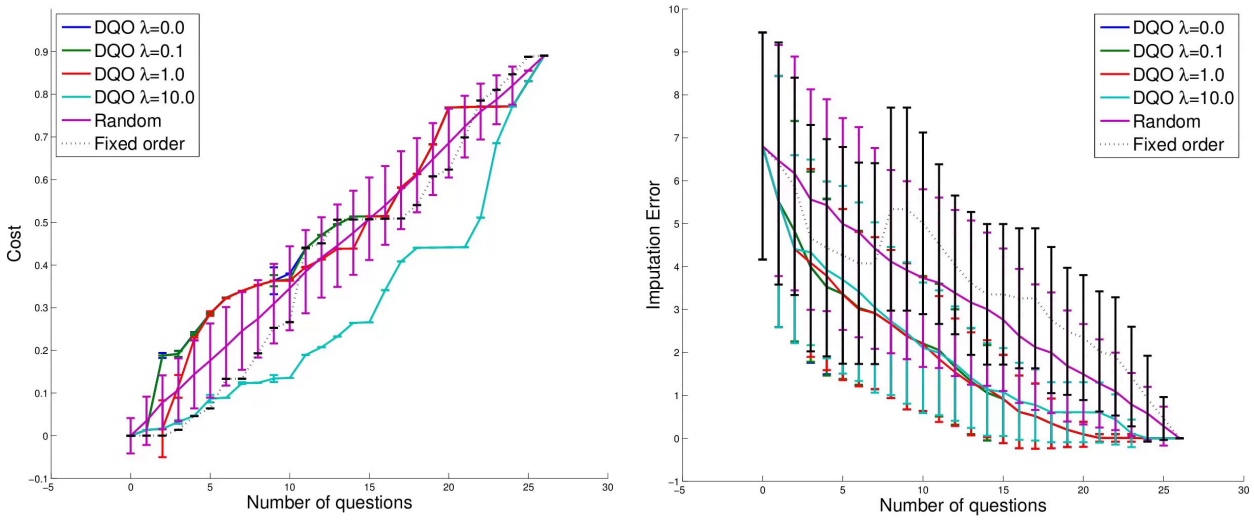
The SIPP is a longitudinal survey, with respondents being interviewed over a period of four years. The previous section considered only a single time point in the interview. In this section, we look at ordering questions asked of a respondent throughout their entire participation in the SIPP.



(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed-order baseline.

(b) DQO has lower imputation error than the fixed-order baseline.

Figure 6.2: Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed-order baseline (dashed lines).



(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed-order baseline.

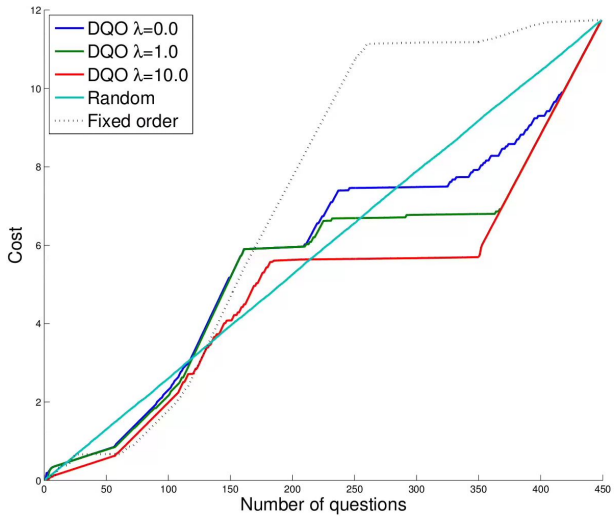
(b) DQO has lower imputation error than the fixed-order baseline.

Figure 6.3: (Same plot as Figure 6.2, but with error bars) Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed-order baseline (dashed lines).

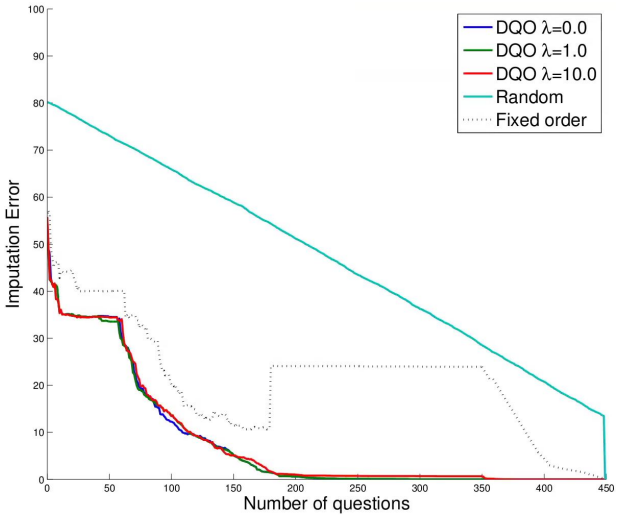
Simulating breakoff in survey collection

Next, we simulated survey breakoff at the presentation of each question by randomly choosing if a respondent would "break off" from the survey, with probability according to the cost of the current question. If the person did break off, their responses to that question and all remaining questions would be left blank.

Figure 6.5 plots the cost (Fig. 6.5a) and imputation error (Fig. 6.5b) trajectories for DQO with several cost penalties λ, the random ordering for each respondent, and the fixed-order baseline. DQO consistently attains lower imputation error than the fixed-order baseline and the random ordering. The imputation error is similar for all DQO orderings, regardless of the cost penalty. The cost penalty does affect the cost trajectory,



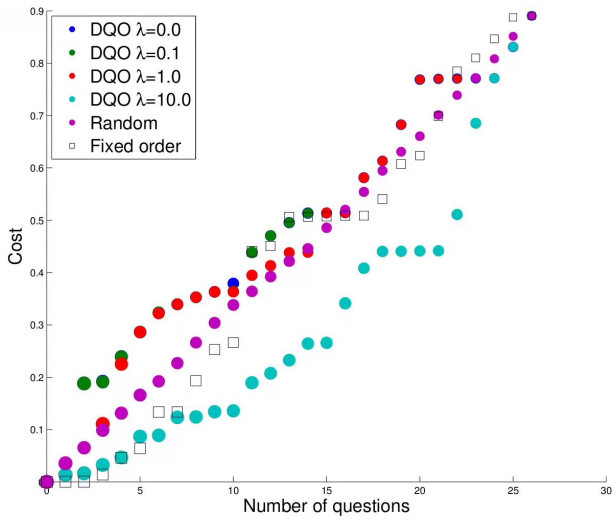
(a) DQO is similarly or less costly than the fixed-order baseline.



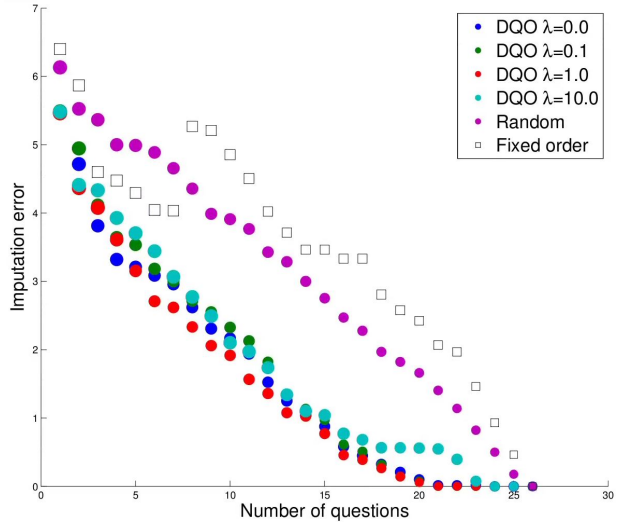
(b) DQO has lower imputation error than the fixed-order baseline.

Figure 6.4: Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed-order baseline (dashed lines).

with higher cost penalties resulting in lower cost trajectories. Figure 6.6 re-plots these results, with more focus given to breakoff rates when each number of questions has been asked: the color of line segments in the plot represents when at least 75% (green), 50-75% (blue), and fewer than 50% (red) of respondents have answered that number of questions. DQO consistently attains lower imputation error than the fixed-order baseline and the random ordering. The imputation error is similar for all DQO orderings, regardless of cost penalty. The cost penalty does affect the cost trajectory, with higher cost penalties resulting in lower cost trajectories.

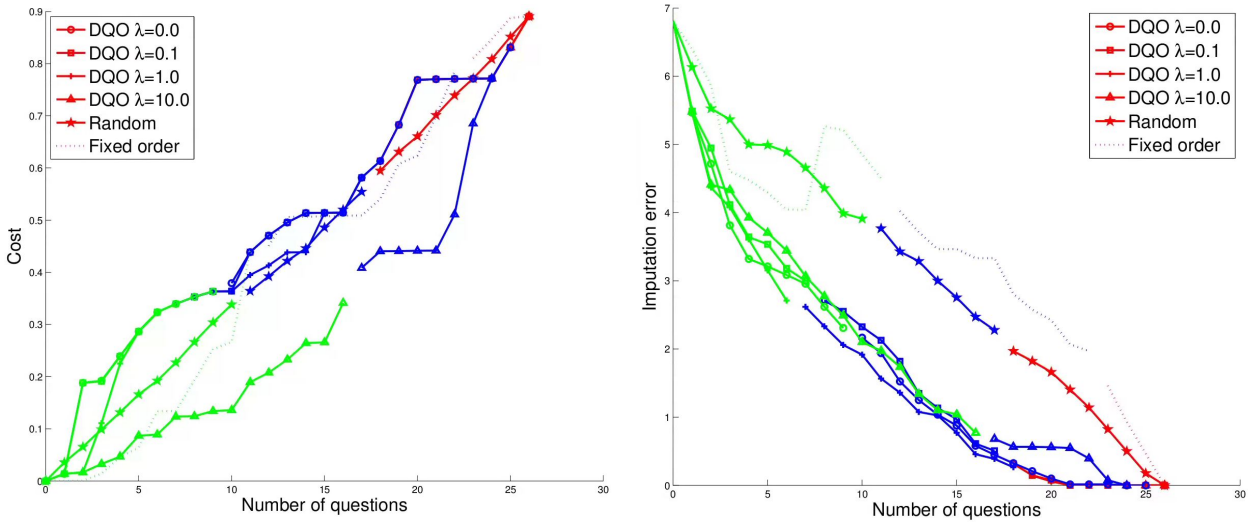


(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed-order baseline.



(b) DQO has lower imputation error than the fixed-order baseline.

Figure 6.5: Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (circles) and the fixed-order baseline (squares). The size of each marker reflects how many respondents reached that far in the survey.



(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed-order baseline.

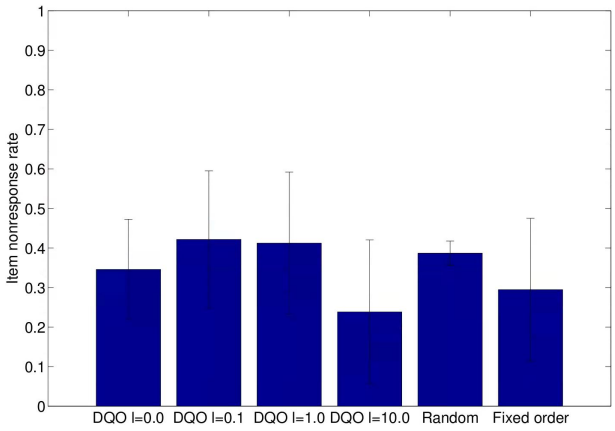
(b) DQO has lower imputation error than the fixed-order baseline.

Figure 6.6: Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed-order baseline (dashed line). The color of each segment reflects how many respondents reached that far in the survey: at least 75% answered questions in green, 50-75% answered questions in blue, and fewer than 50% answered questions in red.

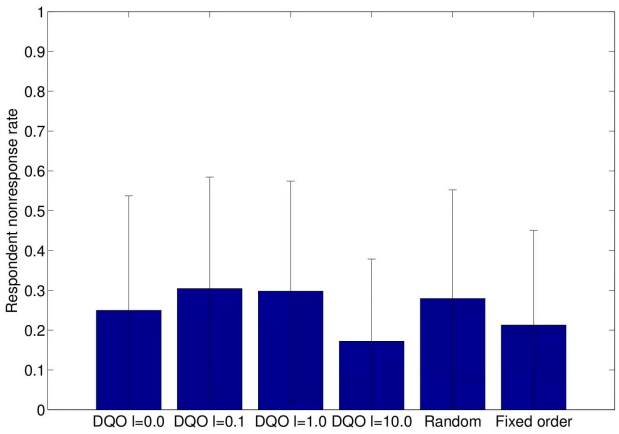
Figure 6.7 illustrates the nonresponse rates for the different methods: Figure 6.7a plots the average nonresponse rate for each question (fraction of respondents who answered each question before dropping out), and Figure 6.7b plots the average nonresponse rate for each respondent (fraction of questions each respondent answered before dropping out). For both types of nonresponse, DQO performs similarly to the fixed-order baseline when $\lambda = 0$, slightly worse when $\lambda = 0.1, 1.0$, and better when $\lambda = 10$. Figure 6.8 shows individual item nonresponse rates, compared to the fixed-order baseline. This plot illustrates how DQO can achieve lower nonresponse (*i.e.*, higher values in the plot of "baseline item nonresponse rate" − "DQO item nonresponse rate") for later items in the survey, since those items might be asked earlier and therefore are less likely to suffer from dropout.

Figure 6.9 plots the final imputation error for each method. Final imputation error is the total number of missing values that were incorrectly imputed (via k-NN) for a respondent once that respondent stops providing answers. For this metric, all variants of DQO outperform the fixed-order baseline. Even though some of the DQO orderings have higher nonresponse rates than the fixed-order baseline (Figure 6.7), DQO tends to choose the most informative questions near the beginning so imputation quality does not suffer even when a respondent drops out before completing the survey.

Bias and variance of survey estimates To compare performance in terms of the quality of final survey estimates (*e.g.*, population means), we set the DQO orderings and the fixed orderings to have 16% breakoff rates (the average breakoff rate for web surveys (Manfreda & Vehovar, 2002)). We ran multiple imputation to get multiple sets of imputations for the values missing due to dropout (Rubin, 2004), using the SAS multiple imputation procedure with a fully conditional specification (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) to model each variable to be imputed. The k-NN-based imputation error metric used earlier in this section measured how closely imputed values for yet-unanswered questions from one individual respondent matched the true values for that respondent. In contrast, this multiple imputation approach captures the uncertainty associated with each variable (across the multiple sets of imputations) and therefore can gives better estimates at the population level. After multiply imputing values for the dropout cases, we calculated the bias and variance of the resulting survey estimates. Table 6.1 shows that the estimates from DQO have bias and variance lower than or similar to those from the standard SIPP order. However, the differences between the DQO and SIPP orders are not statistically significantly different.



(a) The cost penalty λ affects the item nonresponse rates in DQO.



(b) The cost penalty λ also affects the respondent nonresponse rates in DQO.

Figure 6.7: Plots showing nonresponse rates for each question (fraction of respondents who did not answer a question, due to dropout) and each respondent (fraction of questions each respondent did not answer when they broke off): lower values are better.

Table 6.1: Bias and variance of final survey estimates for SIPP order and DQO orders. Bias was calculated as the (absolute value of the) difference between the calculated survey estimates (which include imputed values) and the true values (from the complete data). Variance was calculated as the variance of survey estimates, across the multiply-imputed datasets. For both metrics, means and standard errors are given across the survey variables.

Method	**Bias**	**Variance**
SIPP order	0.063 ± 0.194	0.013 ± 0.056
DQO, $\lambda = 0$	0.034 ± 0.111	0.005 ± 0.023
DQO, $\lambda = 0.1$	0.047 ± 0.103	0.007 ± 0.031
DQO, $\lambda = 1.0$	0.051 ± 0.146	0.012 ± 0.054
DQO, $\lambda = 10$	0.024 ± 0.063	0.001 ± 0.004

6.2.4 Limitations

The SIPP Synthetic Beta (SSB) dataset has several downsides. First, household structure is not preserved—each respondent is treated as a separate entity. Though there are spouse links for female-male married couples, this information is insufficient to recover a household structure which may include same-sex married couples, unmarried partners, children, other relatives, and other household members. Second, because the SSB is synthetic and fully imputed, it does not have any information on when respondents break off from the survey. However, since we simulated breakoff as proportional to question cost (measured as item nonresponse rate), we expect our breakoff-related results to hold for the case of known breakoff when breakoff probability is the cost metric.

6.2.5 Conclusion

We can capitalize on the relationships among survey responses when dynamically ordering questions to collect the most informative data first, so that if a respondent drops out before finishing the survey, imputation quality will be good. Additionally, when such a method for dynamic question ordering takes breakoff probability into account when ordering questions, breakoffs will be delayed until after the most relevant information is collected. Although this approach means that certain items are more likely to be unanswered (*i.e.*, high-cost items), the dynamic process collects enough low-cost information from the respondent before

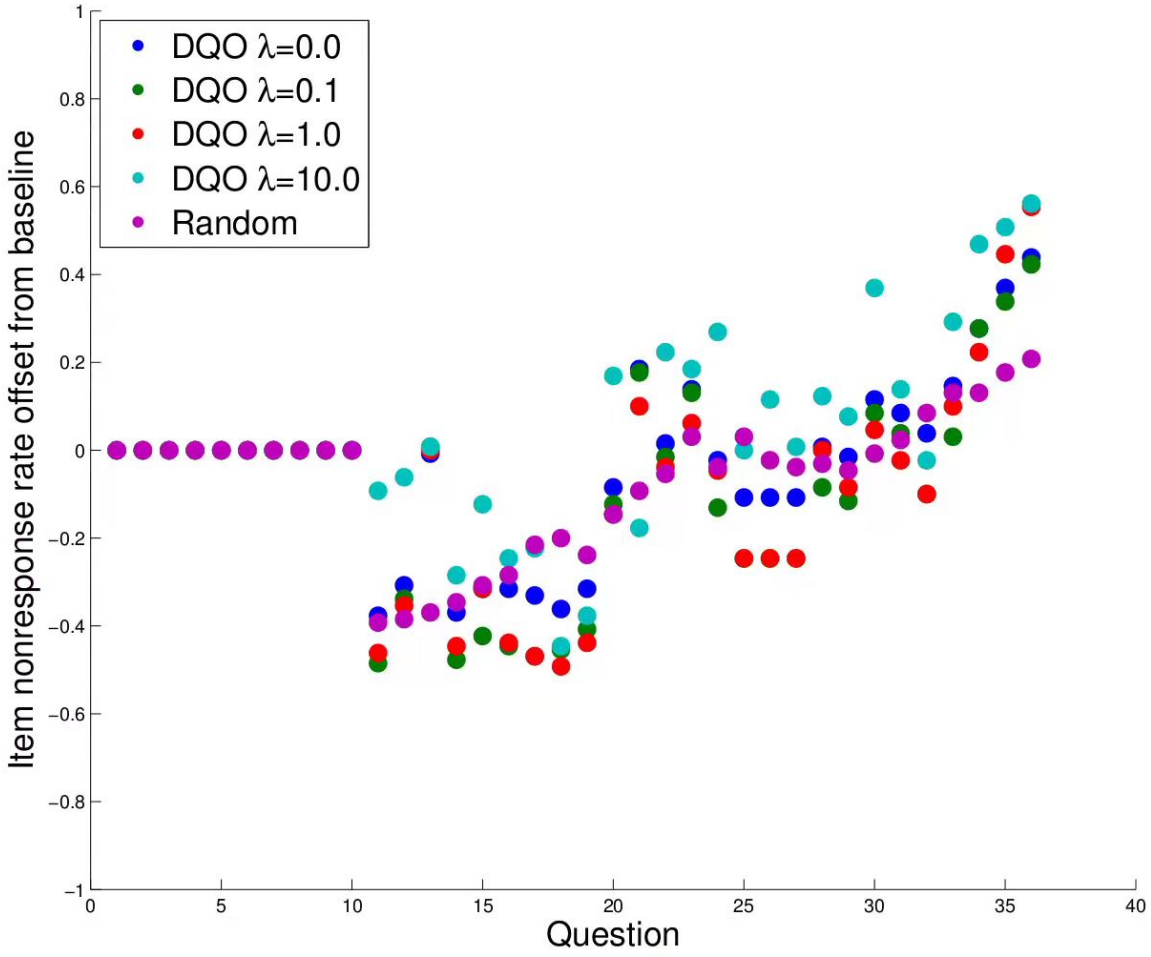


Figure 6.8: The offset of item nonresponse rates from the baseline's: positive values mean a method has lower item nonresponse rate than the fixed-order baseline. The first ten items come from administrative records and are therefore "free" (they are not asked during the interview).

they break off, so that imputation quality is higher than a fixed order for the questions. Furthermore, even when DQO and standard question orders are fixed to have the same amount of breakoff, DQO collects information such that final survey estimates have bias and variance that are less than or similar to those of estimates from a fixed question order.

Simulated breakoff on the SSB, in Section 6.2.3, shows the potential of DQO to gather relevant information from respondents prior to breakoff. The next step is to work with survey data that records actual instances of breakoff from respondents, which we do in our next application on the American Community Survey.

6.3 Application to the American Community Survey

For the mandatory American Community Survey (ACS), the goal is to gather complete statistics on the U.S. population, and follow-up with nonrespondents is expensive. Each year 3.54 million households receive mailed surveys to answer anywhere between 77 and 347 questions, depending on the number of household occupants (U.S. Bureau of the Census, 2016). The survey takes, on average, 40 minutes to complete and 54% of homes return theirs (U.S. Bureau of the Census, 2014a). The Census Bureau calls nonrespondents for telephone interviews and then samples nonrespondents for home interviews. Each in-person case takes 134 minutes, and in 2012 this amounted to 129,000 person-hours per month (Griffin & Nelson, 2014). In addition to being expensive, in-person interviews can bias survey results, if nonresponse adjustment weights

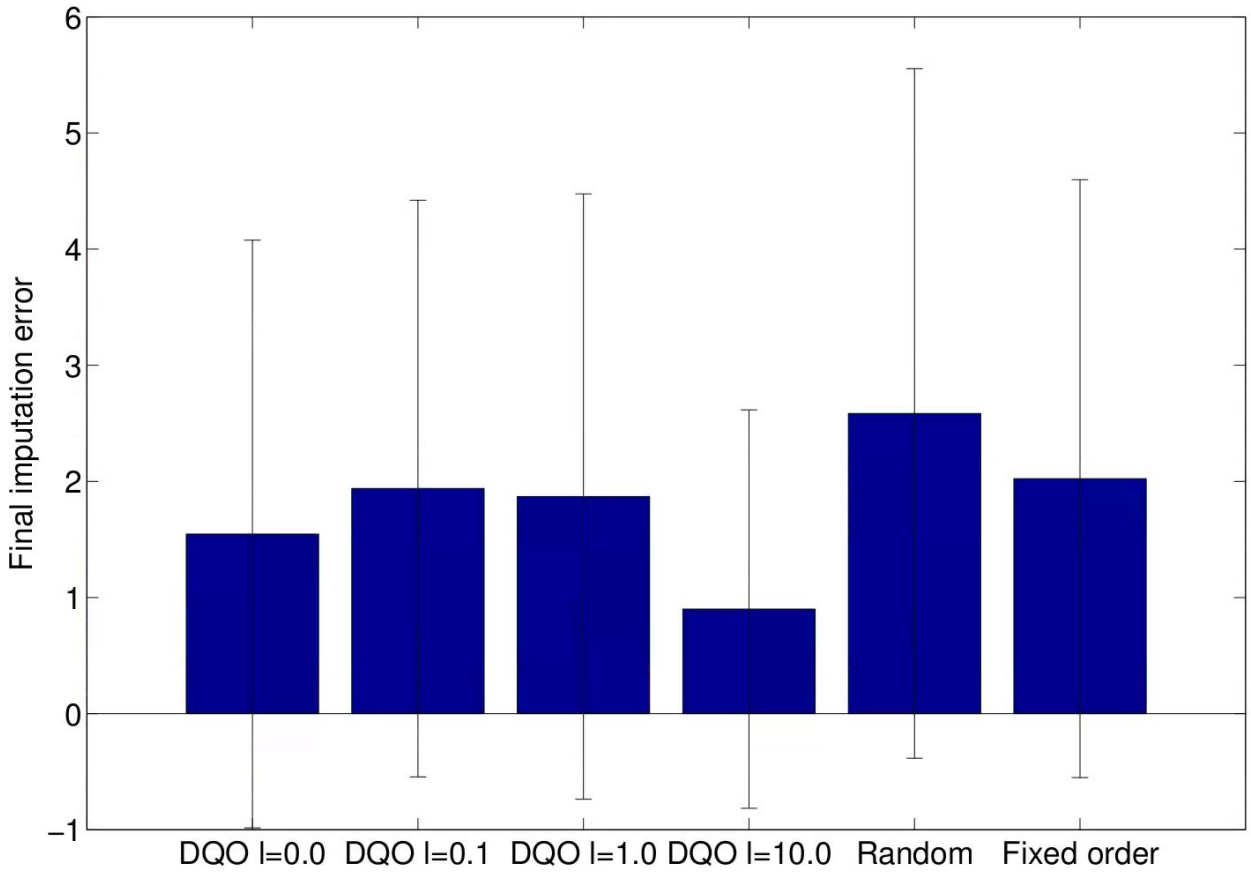


Figure 6.9: Final imputation error (number of incorrectly imputed values once the respondent finishes (including by dropping out)) the survey: lower values are better.

are not allocated correctly (U.S. Bureau of the Census, 2014a).

The Census Bureau tested shifting the mail survey online and found similar data quality for Internet and mail return (Horwitz et al., 2012). Furthermore, while overall response rates were similar, online surveys had higher item response rates for earlier questions and more blank responses for later questions than paper surveys (Horwitz et al., 2012). Dynamically ordering survey questions in the online form could ensure that even if households do not complete the survey, they answer the most informative questions before breaking off.

The online mode for the ACS also collects paradata as respondents complete the survey. These paradata include clicked links (including navigation buttons, responses, help buttons), timestamps, field values, errors, invalid logins, timeouts, logouts (Horwitz et al., 2012). Such paradata could be used to model user engagement, understanding, and willingness to respond, as another component for dynamic question ordering to increase response rate.

6.3.1 The ACS online mode

The Internet questionnaire for the ACS has four main sections to collect information from a respondent. The first section of the survey establishes the *roster* for the household: which people live at the sampled address. The survey first asks the respondent to list the people who live or stay at their address. Several subsequent questions prompt the respondent to add forgotten names or remove additional names (*e.g.*, people who are away for more than two months, such as a college student). Once the roster of relevant inhabitants has been finalized, the second section of the survey asks about who owns or rents the home, the relationships of the other household members to this reference person, and demographic information on all inhabitants

(sex, date of birth, Hispanic origin, and race). Next, the third section of the survey collects information about the housing unit, such as the year it was built, how much electricity cost in the past month, and if a household member has a mortgage on the home. Finally, the fourth section asks detailed questions for each member of the household, such as place of birth, income in the past year, and the length of their commute from home to work. These detailed person questions are asked for one occupant, then the next, *etc.* Between each person, there is a "person select" ("pselect") screen on which the respondent can select for which household occupant they will answer the questions next[2]. Once a respondent has completed all the household questions and detailed person questions for each household occupant, they have the opportunity to review (and possibly edit) their answers before submitting the survey.

If a respondent wishes to complete the survey at a later time, they can save their answers and log out. In order to log in again, they need to enter a four-digit code they received upon their first login to the survey. At the time that the data used in this chapter were collected, it was not possible to reset the code if the respondent forgot or lost it. Therefore, respondents who logged out before completing the survey and later tried to log in without their code received a "failed login" error and could not access the survey again. The current implementation of the online ACS prompts respondents to answer a security question when they first log in to the survey so that they can recover access to the survey if they log in later without their code.

Paradata in the online ACS

In addition to collecting respondents' answers to survey questions, the online mode of the ACS also passively collects information about their progress in the survey, *paradata*. These paradata log events in the survey process: logins, logouts, field changes, hyperlink clicks, *etc.* Each event has a timestamp, a household identifier, the page of the survey on which it occurred, the text for any error messages that occurred, and information about a respondent's device (type of device, operating system, browser) when they logged in to the survey.

We have response data and paradata from the online respondents in the June through September 2016 panels of the ACS[3].

6.3.2 Breakoff in the online ACS

A response is considered "complete" if the respondent reaches at least the presummary screen (the page where they can review their answers). If a respondent stops answering questions before this point (and therefore does not answer all the questions they are meant to answer), their response is considered a "breakoff." Of the 325,000 respondents in our four-month dataset, 36,630 (11.27%) broke off before completing the entire questionnaire.

Related work on breakoff in online surveys

Past work has identified factors that can influence breakoff in survey design, page and question characteristics, and respondent characteristics (Peytchev, 2009). Features of the survey design, such as a progress indicator that shows respondents how much of the survey they have completed so far, also affect breakoff. When progress is shown early in the survey, respondents are more likely to break off (Peytchev, 2009). "Encouraging" feedback (*i.e.*, feedback that the respondent is proceeding quickly through the survey) makes respondents less likely to break off (Conrad, Couper, Tourangeau, & Peytchev, 2010). Page and question characteristics that are related to breakoff include open questions and long questions, which increase the risk of breakoff, as well as pages that begin a new section of the survey (Peytchev, 2009).

Finally, demographic characteristics of respondents can be associated with survey breakoff. For example, members of these groups disproportionately break off from surveys: people with low education levels (a proxy for cognitive sophistication) (Peytchev, 2009), people of color (Peytchev, 2011), men (Peytchev, 2011), and younger respondents ("younger" as defined by age (Peytchev, 2009; McCutcheon, al Baghal, & Tsabutashvili, 2013) and by seniority, for college students (Peytchev, 2011)).

[2]The ACS is designed so that one person may fill it out for the household if they know all of the requested information, but they also can consult with other household members or have multiple people in the household complete parts of it.

[3]Because this is a partial dataset (online self-respondents from four months of ACS data collection), it is not possible to assess the external validity of these results using the standard Census approach of replicate weights.

Additionally, there has been evidence that people who are likely to break off from a survey are also more likely to be nonrespondents, and *vice versa* (Peytchev, 2011). Thus, the following demographic groups that are disproportionately nonrespondents may also be more likely to break off from surveys: low-income people (Goyder, Warriner, & Miller, 2002; Van Goor & Rispens, 2004), elderly people (Kaldenberg, Koenig, & Becker, 1994), people of color (Johnson, O'Rourke, Burris, & Owens, 2002; Groves et al., 2009; Peytchev, 2011; Keeter, Kennedy, Dimock, Best, & Craighill, 2006), men (Peytchev, 2011; Gray, Campanelli, Deepchand, & Prescott-Clarke, 1996), and people with low education levels (Hauser, 2005; Keeter et al., 2006).

Furthermore, breakoff and response time are tightly connected. Peytchev (2009) found that "respondent behavior as time spent on the first question was significantly associated with risk of breakoff, with each standard deviation slower responding resulting in 34% higher risk of breakoff." Thus, factors that influence response time can also influence breakoff rates. For example, question length (Bassili & Scott, 1996; Yan & Tourangeau, 2008), question complexity (Wagner-Menghin, 2002; Yan & Tourangeau, 2008), and drop-down menus (Healey, 2007) increase item response times. Respondents who complete an online survey on a smartphone can take more time to complete the survey and are more likely to break off than respondents who used a computer (Lambert & Miller, 2015).

6.3.3 Results on the ACS

First, we examine differences between the populations of people who broke off compared to those who completed the survey. Then, we look at characteristics of questionnaire pages and their breakoff rates. Next, we analyze how respondent characteristics, page characteristics, and respondent behavior influence breakoff page-by-page. Finally, we experiment with adaptively ordering questions for individuals, to reduce and delay breakoff.

Respondent characteristics and their effect on breakoff

The ACS form grows in length with the number of household members, since the detailed person questions are asked for each occupant. Furthermore, complex household structures involving, for example, unrelated people such as roommates or boarders increase the difficulty for the respondent to complete the form. Table 6.2 shows the number and percentages of breakoffs and complete responses for households of various sizes. Breakoff rates increase as household size, and therefore questionnaire length, increase. Table 6.3 shows breakoff and completion rates for the most prevalent household structures (collections of individuals and their relationships to one another) in the ACS dataset. The breakoff rates by household structure are influenced by the household size (larger household structures have greater breakoff), as well as the relationships between household members. For example, the "reference + spouse" and "reference + unmarried partner" household structures both contain two people, but "reference + spouse" has a 7.61% breakoff rate, compared to the 11.58% breakoff rate of "reference + unmarried partner." Presumably spouses know one another better than unmarried partners, which could explain the higher completion rate in households containing only spouses compared to those containing only unmarried partners. Table 6.4 shows the breakoff counts for households with and without nonrelatives living together. Most households contain only relatives living together, but the 4.85% of households containing nonrelatives have a 25.92% breakoff rate, compared to the 10.52% breakoff rate for those households containing only related people.

Device usage also differed across the completion conditions. Each time a respondent logs in to the online ACS instrument, information about the device the respondent is using to access the website is logged in the paradata. Table 6.5 shows how many respondents logged in to the survey instrument one, two, and three or more times, as well as the number of respondents whose logins were not recorded. 5.17% of respondents did not have their logins recorded properly, and these respondents broke off from the survey at a higher rate (27.38%) than the respondents whose logins were recorded correctly (10.39% breakoff). Perhaps the technical issues that caused these respondents' logins not to be recorded properly also caused technical glitches to the survey instrument that made it challenging for these respondents to complete the survey. Tables 6.6a and 6.6b focus on the respondents whose logins were recorded correctly and show how frequently respondents used computers (desktops and laptops), mobile devices (smartphones, tablets, e-readers, *etc.*), and other devices (PlayStations, smart TVs, *etc.*) to access the online ACS form. The most prevalent device was computers, then mobile devices—80.13% of single-session respondents used computers, and 19.86% used mobile devices.

Table 6.2: Complex household structures, with multiple people, have higher incidences of breakoff than simple household structures.

Number of household members	Breakoff	Complete
1	3426 (4.92)	66210 (95.08)
2	11180 (9.05)	112400 (90.95)
3	7630 (14.20)	46110 (85.80)
4	7428 (15.85)	39440 (84.15)
5	3983 (19.78)	16150 (80.22)
6	1638 (23.33)	5383 (76.67)
7+	1351 (32.71)	2779 (67.29)
Total	36630 (11.27)	288400 (88.73)

Table 6.3: Complex household structures (multiple people, unrelated people such as roommates or boarders, *etc.*) have higher incidences of breakoff than simple household structures. This table shows the most frequent household structures and their breakoff and completion rates.

Household structure	Breakoff	Complete
Reference + spouse	6764 (7.61)	82090 (92.39)
Reference + spouse + child(ren)	12000 (13.63)	76030 (86.37)
Reference	3741 (5.34)	66280 (94.66)
Reference + child(ren)	2689 (12.63)	18610 (87.37)
Reference + unmarried partner	1410 (11.58)	10770 (88.42)
Reference + nonrelative(s)	2091 (24.31)	6510 (75.69)
Others	7936 (21.99)	28150 (78.01)
Total	36630 (11.27)	288400 (88.73)

Mobile respondents broke off at higher rates than computer respondents, in agreement with past work (*e.g.*, (Lambert & Miller, 2015))—15.39% of single-session mobile respondents broke off, compared to 9.53% of single-session computer respondents. For respondents who spent multiple sessions on the survey, we look at the first and last devices they used to access the survey in Table 6.6b. Most multi-session respondents used the same type of device for their first and final logins (80.68% used computers first and last, and 14.33% used mobile devices first and last), and those who switched devices were more likely to switch from a mobile device to a computer than *vice versa* (1.96% of respondents whose first login was on a computer had their final login on a mobile device, and 19.02% of respondents whose first login was on a mobile device had their final login on a computer). Breakoff rates were lowest for respondents whose first and final sessions were on a computer.

Breakoff patterns by respondent demographic group in the online ACS agree with most trends found in previous work. Tables of breakoff and completion status by these demographic groups appear in Appendix C. One demographic characteristic of breakoffs that was not reproduced in this ACS dataset is the past finding that men are more likely to break off than women (Peytchev, 2011). Table C.1 shows the numbers of breakoff and complete responses from women and men—slightly more women than men were respondents for the survey (52.20% of respondents were women), and women broke off at a higher rate than men (12.37% breakoff for women, 9.78% breakoff for men). Remaining demographics and breakoff were observed in accordance with prior work. Table C.2 shows that older respondents completed the survey at higher rates than younger respondents (McCutcheon et al., 2013; Peytchev, 2009, 2011). Tables C.4 and C.3 show that nonwhite and Hispanic respondents, respectively, were more likely to break off than white and non-Hispanic respondents (Peytchev, 2011). Table C.5 shows that respondents with higher education levels completed the survey at greater rates than respondents with lower education levels (Peytchev, 2009). Finally, Table C.6 shows that low-income respondents were more likely to break off than higher-income respondents (Peytchev,

Table 6.4: Complex household structures (multiple people, unrelated people such as roommates or boarders, *etc.*) have higher incidences of breakoff than simple household structures. This table shows the completion status for households with and without nonrelatives.

Household structure	**Breakoff**	**Complete**
No unrelated household members	32540 (10.52)	276700 (89.48)
Unrelated household members	4089 (25.92)	11680 (74.08)
Total	36630 (11.27)	288400 (88.73)

Table 6.5: Number of logins to the online ACS, by breakoff or complete status. Percentages of each login category that are breakoffs or completes are in parentheses. More people who break off do so after one login than multiple logins.

Number of sessions	**Breakoff**	**Complete**
Missing	4604 (27.38)	12210 (72.62)
1	27170 (10.69)	226900 (89.31)
2	2749 (8.59)	29240 (91.41)
3+	2112 (9.51)	20100 (90.49)
Total	36630 (11.27)	288400 (88.73)

2009; Goyder et al., 2002; Van Goor & Rispens, 2004).

In our analyses of breakoff from the online ACS, we do not include these demographic variables because the respondent does not provide this information until partway through the questionnaire. Thus, this information is unknown until a respondent reaches these questions (and always unknown, for a large portion of respondents who break off).

We can use paradata collected from the online ACS form to track how frequently respondents changed their answers. Table 6.7 shows completion rates by number of total field changes respondents made. Most respondents made zero or few field changes: 37.92% made no field changes, 26.28% made one, 16.34% made two, 16.36% made three through five, and only 3.11% made six or more field changes. Higher field change counts (up to a point) are associated with higher completion rates: 15.61% of respondents with no field changes broke off, and this breakoff percentage decreased down to 5.82% among respondents who made five field changes. This observation is partly due to the fact that people who complete the survey typically visit more pages than people who break off and therefore have more opportunities to make field changes. However, a high number of field changes can indicate confusion with the survey questions or instructions and therefore precede breakoff (*e.g.*, 40.96% of the respondents with over 15 field changes broke off).

Page characteristics and their effect on breakoff

Next, we consider characteristics of pages and questions that might influence breakoff. Figure 6.10 shows the fraction of respondents who reached the n-th page in the online ACS questionnaire. The fractions of respondents are scaled using the number of people who should have reached a page, based on the number of people living in the household. Due to more complex screening rules, the universe for some pages changes and thus those pages are underestimated in terms of fraction (*e.g.*, the question about births in the past year is asked only about women age 15 to 50). Table 6.8 lists the top ten pages on which respondents broke off.

Most pages in the online ACS questionnaire display a single multiple-choice question, which is answered with radio buttons or checkboxes (Figure 6.11). Some pages have more complicated questions or formats, which can affect respondent behavior. We include variables to indicate when pages have open questions (Figure 6.12), questions with drop-down menus for answer choices (Figure 6.13), multiple questions (Figure 6.14), or a question that indicates that a new section is approaching (Figure 6.15), all of which have been shown to influence breakoff (*e.g.*, (Peytchev, 2009)). Eight of the top ten breakoff pages in Table 6.8 fall into

Table 6.6: Device category used to complete the online ACS form, by breakoff or complete status, for single- and multi-session survey behavior. The most popular device type was a computer, and computer respondents had the lowest breakoff rates.

(a) Single-session respondents.

Device type	**Breakoff**	**Complete**
Computer	19400 (9.53)	184200 (90.47)
Mobile	7767 (15.39)	42690 (84.61)
Other	1 (5.00)	19 (95.00)
Total	27170 (10.69)	226900 (89.31)

(b) Multi-session respondents.

First-Last device	**Breakoff**	**Complete**
Computer-computer	3664 (8.38)	40070 (91.62)
Computer-mobile	92 (10.54)	781 (89.46)
Mobile-computer	177 (9.70)	1647 (90.30)
Mobile-mobile	927 (11.93)	6841 (88.07)
Other	1 (20.00)	4 (80.00)
Total	4861 (8.97)	49340 (91.03)

at least one category of these more complicated question types: “pselect,” “wages,” and “employeetype” all indicate an upcoming section of the survey, “worklocal” and “wagesamt” have open questions, “worklocal” and “dateofbirth” have multiple questions on their pages, and “dateofbirth” and “placeofbirth” have answer options presented on drop-down menus.

Table 6.7: Counts of respondent field changes by completion type. Percentages of each field change count category that are breakoffs or completes are in parentheses. Many people who broke off made no field changes before breaking off. Increasing field changes are associated with higher completion rates, up until a respondent makes ten or more field changes.

Number of field changes	**Breakoff**	**Complete**
0	19230 (15.61)	104000 (84.39)
1	8801 (10.30)	76620 (89.70)
2	4258 (8.02)	48860 (91.98)
3	2064 (7.05)	27190 (92.95)
4	1002 (6.38)	14700 (93.62)
5	479 (5.82)	7751 (94.18)
6-10	571 (6.30)	8495 (93.70)
11-15	102 (13.84)	635 (86.16)
>15	120 (40.96)	173 (59.04)
Total	36630 (11.27)	288400 (88.73)

Table 6.8: Top 10 pages on which respondents broke off on the online ACS.

Page	**Number of breakoffs**	**Percentage of breakoffs**
pselect	4502	12.29
worklocal	2349	6.413
wagesamt	1991	5.435
dateofbirth	1898	5.181
placeofbirth	1704	4.652
wages	1301	3.552
vrfyincome	1102	3.008
worklastweek	968	2.643
employeetype	759	2.072
insurance	720	1.966
All others	19340	52.79
Total	36630	100

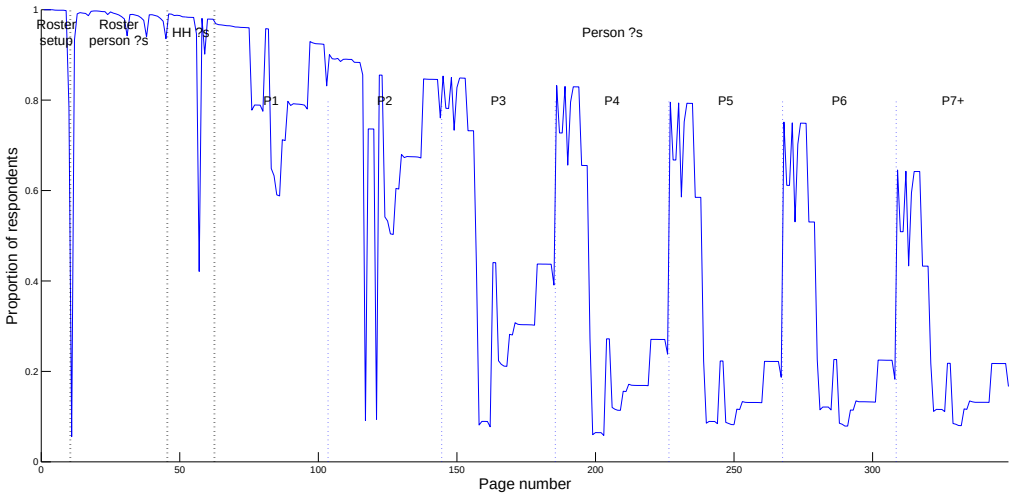


Figure 6.10: Fraction of respondents who reached each page in the online ACS survey. The major sections of the survey are indicated with thick vertical dashed lines—the roster setup (when a respondent indicates how many people live in the household and who they are), the person questions on the roster section (when the respondent gives basic demographics, such as sex and date of birth, about each person on the roster), the household questions ("HH ?s"), and the detailed person questions ("Person ?s"). Thinner vertical dashed lines in the person questions section indicate which person questions are in that section.

Figure 6.11: Most pages in the online ACS have one multiple-choice question, such as this question that asks whether a household member speaks a language other than English at home.

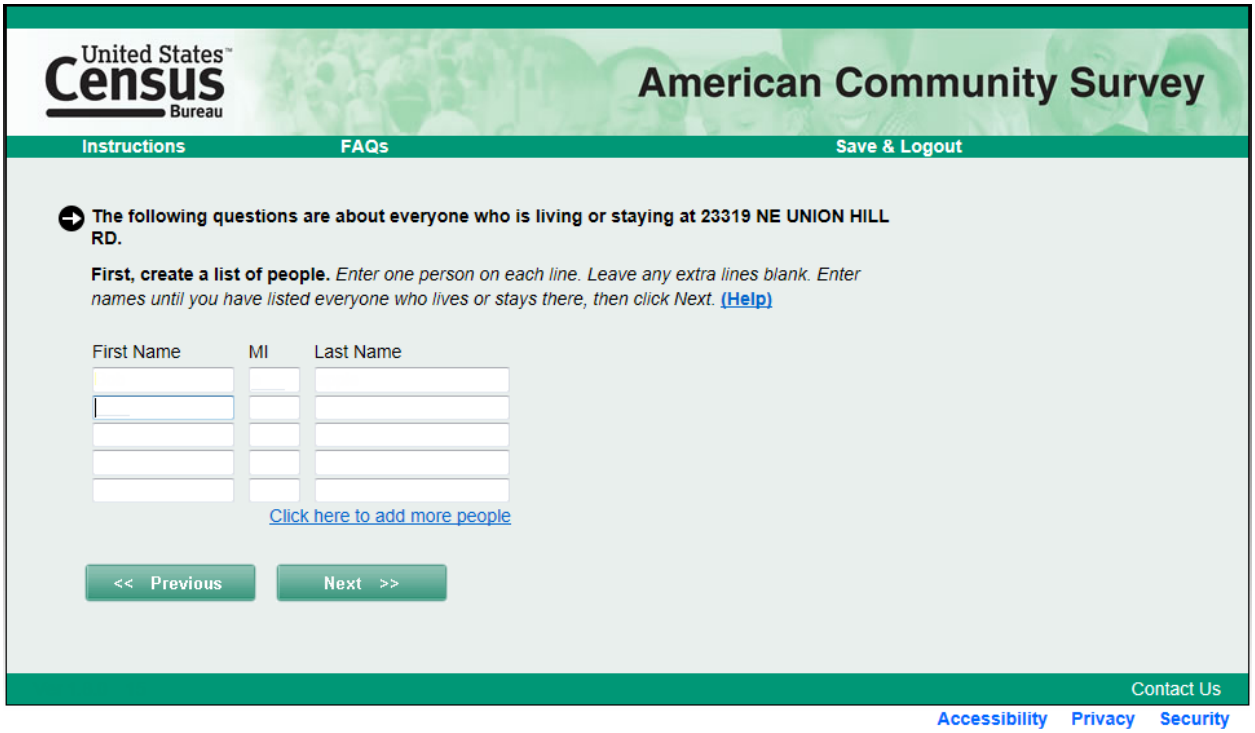


Figure 6.12: Some questions ask respondents to enter information without answer choices, such as this open question about occupants in the household.

Figure 6.13: Some questions ask respondents to enter their answer with a drop-down menu, such as this question about a household member's date of birth.

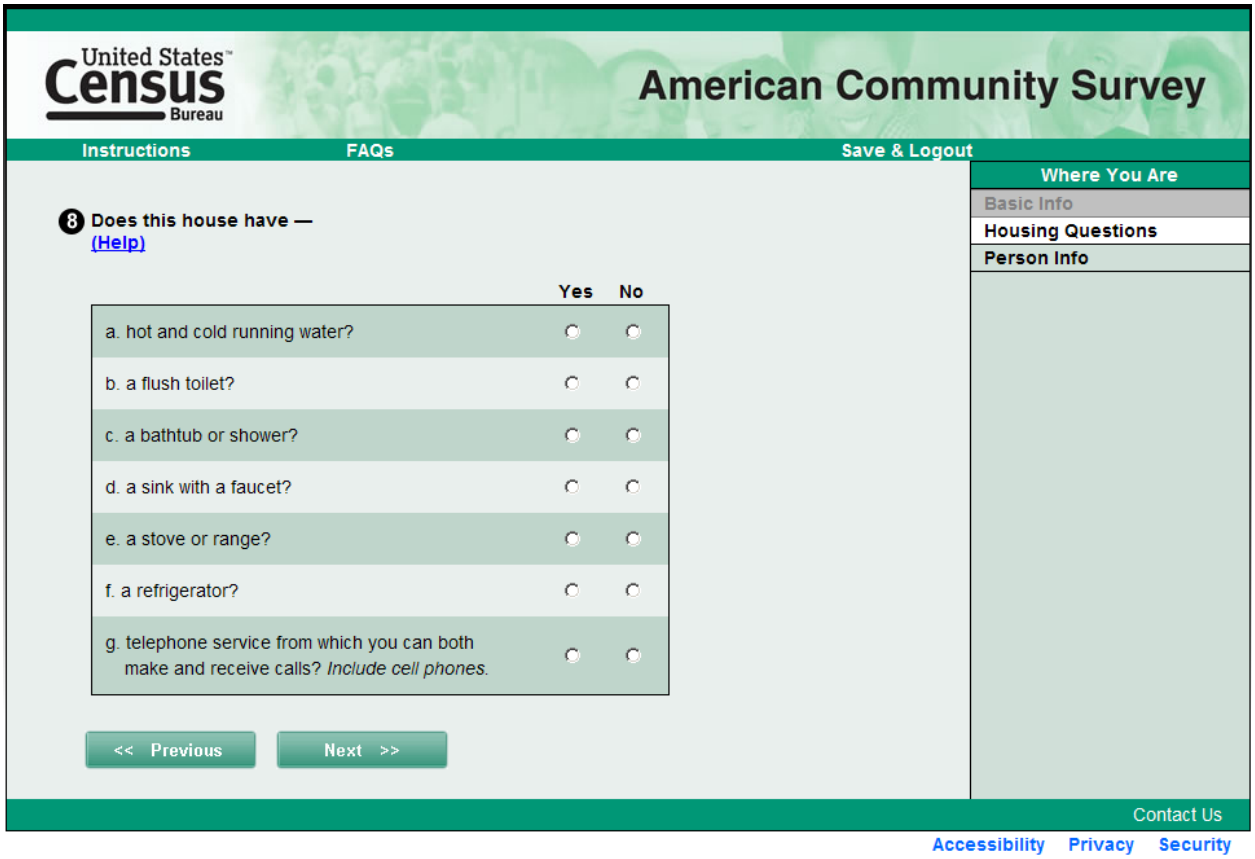


Figure 6.14: Some pages in the ACS include multiple questions, such as this page with seven sub-questions about the facilities available in the residence.

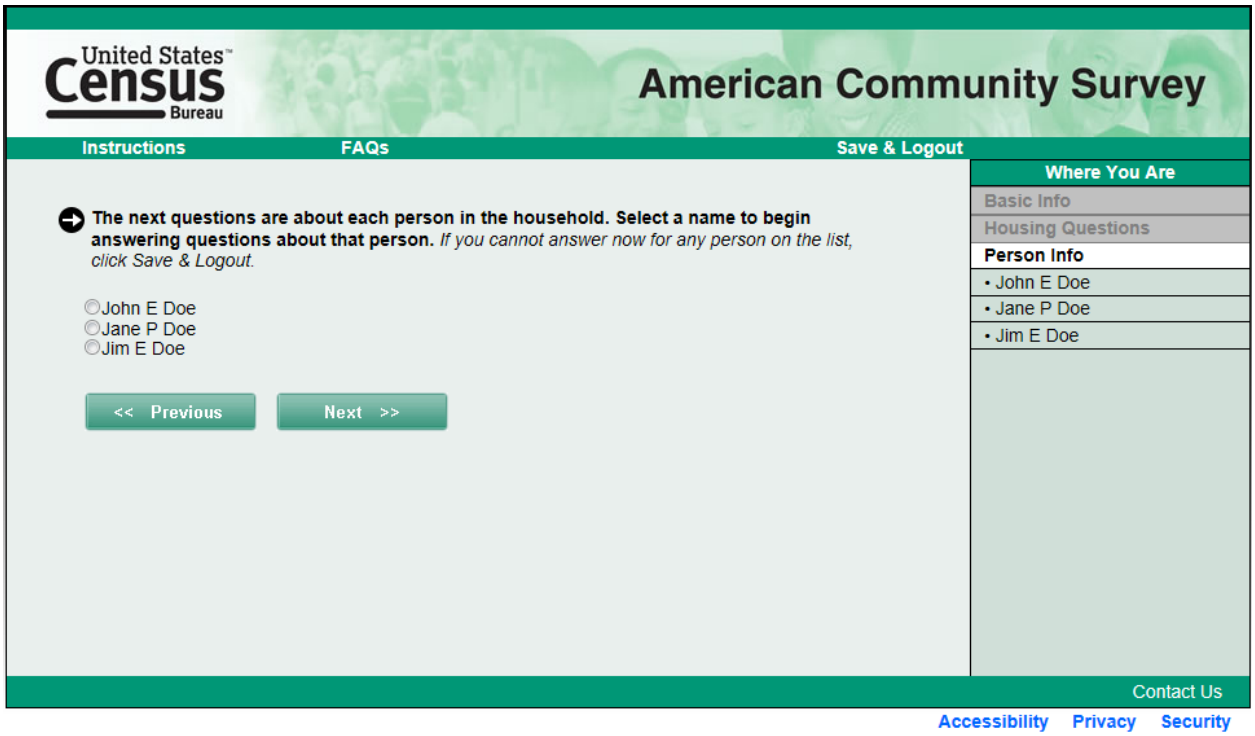


Figure 6.15: Some pages in the ACS alert respondents that a new section of questions is upcoming. For example, this "pselect" question asks a respondent to choose a household member about whom to answer a detailed set of questions next. This page was the most popular page on which respondents broke off from the online ACS (12.29% of breakoffs occurred on the pselect page).

Page-by-page estimates of breakoff probabilities

We use survival analysis (*e.g.*, (Klein & Moeschberger, 2005)) to study when respondents break off from the survey and the factors that influence breakoff. Because the event we observe is discrete (the page a respondent breaks off on), we use a discrete hazard survival model (Cox, 1972; Allison, 1982) for this task.

We follow the approach of Peytchev (2009) to analyze breakoff, including both respondent-specific variables and page-specific variables:

$$\ln\left(\frac{P_{iq}}{1-P_{iq}}\right) = \alpha_q + \beta^T X_i(q), \tag{6.3}$$

where P_{iq} is the probability that respondent i breaks off on page q, α_q is the baseline hazard of breaking off on page q, β is a vector of weights, and $X_i(q)$ is a vector of variables for respondent i on page q.

Our model has one page-invariant variable, the number of household members in a respondent's household ("nper")—this value is constant across all survey pages the respondent visits. We also include page-varying, respondent-invariant variables, like characteristics of a page (whether it is in the setup, roster, household, or person section of the survey) and the question(s) on the page (*e.g.*, if the page asks an open question or has multiple questions). Additionally, we include variables that are both respondent-varying and page-varying, like the cumulative number of field changes a respondent has made up to a point ("field_chg_ct") and the offset between their actual time spent on a survey page and the median time that all respondents spent on that page ("pt_offset").

The screening rules for certain questions and sections in the ACS mean that respondents can reach pages at different points in the survey. For example, only respondents who first answer that their home is part of a condominium reach the subsequent page that asks if there is a condominium fee. Therefore, page q in our model indicates the q-th page that a respondent reaches, and not any specific page. The page-varying variables, like page type and question type on the page, capture some specifics about the particular page and questions a respondent sees at page q.

We consider only those respondents with one login[4] and learn two survival analysis models, one for computer respondents (Table 6.9) and one for mobile respondents (Table 6.10). Most trends of how variables influence breakoff hold across the computer and mobile models. A respondent is slightly more likely to break off on a page where they have spent more time than most respondents. Compared to one-person households, respondents answering for multiple-person households are more likely to break off, with larger effects for larger household sizes (the difference between one- and two-person households is not significant). High numbers of field changes indicate a greater likelihood of breakoff than no field changes. Of the question type variables, a question indicating that a new section of the survey is beginning makes the biggest impact on breakoff likelihood—computer and respondents are 50 and 24 times, respectively, more likely to break off on a page with a section question than on a page without a section question. Open questions and drop-down questions are also associated with higher breakoff, for both computer and mobile respondents. Computer respondents are more likely to break off on a page with multiple questions than a page with a single question, but mobile respondents are less likely to break off on a page with multiple questions.

6.3.4 Dynamic question ordering on ACS

We perform DQO on a test set of ACS data, for computer and mobile respondents. For this application of DQO, we use respondent engagement, defined as the probability that a respondent would continue the survey on a particular page, as the utility metric for DQO. We used item response time as the cost metric in DQO. We simulate breakoff by having a respondent break off at a particular point in data collection, according to the predicted breakoff probability calculated from the survival analysis model. For this simulation of breakoff, we restrict our test sets to 1000 respondents who completed the survey on each device type, so we have complete data and can simulate answering questions in any order. We compare completion rates for the dynamically ordered form to those for a fixed-order form that asks questions in the order of the standard ACS questionnaire.

Table 6.11 shows the number of breakoffs for the dynamically ordered forms, with three values of tradeoff parameter λ, and the fixed-order form. For both computer and mobile respondents, DQO resulted in fewer

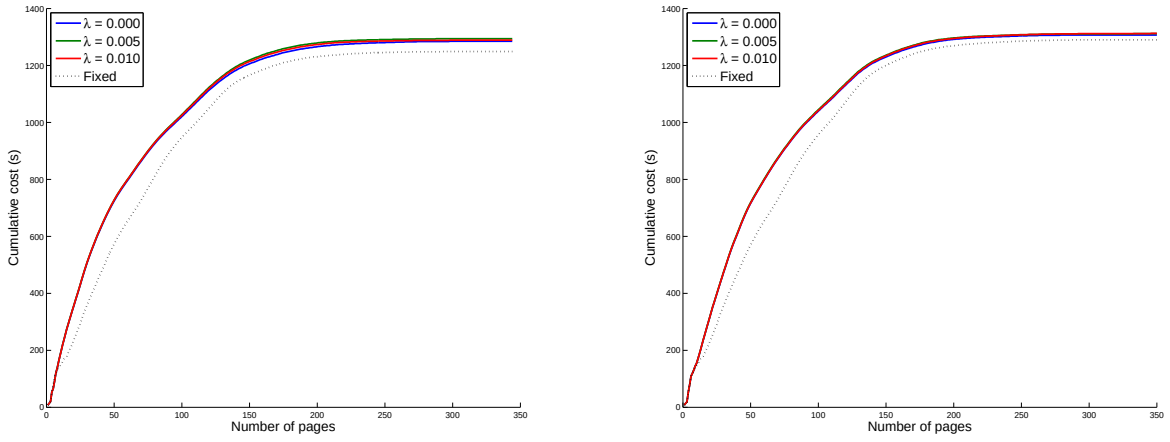
[4]At the time that these data were collected, it was not possible for respondents who logged out from the survey and forgot their login codes to reset their codes or access the survey later to complete it.

Table 6.9: Survival analysis model for computer respondents.

Parameter	Category	β	Standard error	Chi-Square	Pr > ChiSq	Hazard ratio
pt_offset	ctns	0.00287	4.35e-05	4375	<.0001	1.003
nper	2	0.06402	0.03424	3.497	0.0615	1.066
nper	3	0.3967	0.0382	107.9	<.0001	1.487
nper	4	0.4777	0.03991	143.3	<.0001	1.612
nper	5	0.6913	0.04723	214.3	<.0001	1.996
nper	6	0.8242	0.06429	164.3	<.0001	2.28
nper	>6	1.179	0.07386	255	<.0001	3.253
field_chg_ct	1	0.01481	0.02466	0.3607	0.5481	1.015
field_chg_ct	2	0.09235	0.03409	7.34	0.0067	1.097
field_chg_ct	3-4	0.1814	0.04271	18.04	<.0001	1.199
field_chg_ct	5-10	0.6513	0.07172	82.46	<.0001	1.918
field_chg_ct	>10	2.562	0.1578	263.5	<.0001	12.96
page_type	setup	-6.196	0.3704	279.9	<.0001	0.002
page_type	roster_demos	-3.017	0.3504	74.15	<.0001	0.049
page_type	household	-2.473	0.3539	48.85	<.0001	0.084
page_type	person	-2.322	0.3583	41.99	<.0001	0.098
open_q	1	0.9557	0.02218	1857	<.0001	2.601
multi_q	1	0.4971	0.05031	97.65	<.0001	1.644
dropdown_q	1	1.469	0.03894	1423	<.0001	4.344
sec_q	1	3.915	0.03941	9869	<.0001	50.13

breakoffs than the fixed-order form. Additionally, Table 6.12 shows that, on the cases where DQO did still result in breakoff, the breakoffs in the DQO-ordered form happened, on average, several questions later than the breakoff on the fixed-order form.

Figures 6.16a and 6.16b plot the cost trajectories (cumulative time spent on the survey) as successive pages in the questionnaire are reached for computer and mobile respondents, for DQO and fixed orders. In these plots, the DQO orders (solid lines) are higher than the fixed orders (dashed lines) since the reduced and delayed breakoffs for DQO mean that respondents are spending more time on the survey.



(a) Computer respondents.

(b) Mobile respondents.

Figure 6.16: Cumulative time spent on the survey *vs.* page number.

Table 6.10: Survival analysis model for mobile respondents.

Parameter	Category	β	Standard error	Chi-Square	Pr > ChiSq	Hazard ratio
pt_offset	ctns	0.00352	0.0001404	629.8	<.0001	1.004
nper	2	0.01136	0.05075	0.0501	0.8229	1.011
nper	3	0.428	0.05375	63.41	<.0001	1.534
nper	4	0.4957	0.05552	79.7	<.0001	1.642
nper	5	0.7327	0.06179	140.6	<.0001	2.081
nper	6	0.9498	0.07679	153	<.0001	2.585
nper	>6	1.178	0.08801	179.2	<.0001	3.249
field_chg_ct	1	0.00813	0.03307	0.0605	0.8057	1.008
field_chg_ct	2	0.06415	0.04531	2.004	0.1569	1.066
field_chg_ct	3-4	0.2148	0.05366	16.02	<.0001	1.24
field_chg_ct	5-10	0.5017	0.09589	27.38	<.0001	1.652
field_chg_ct	>10	1.697	0.32	28.12	<.0001	5.457
page_type	setup	-6.174	0.7378	70.04	<.0001	0.002
page_type	roster_demos	-2.33	0.7252	10.32	0.0013	0.097
page_type	household	-1.749	0.7293	5.752	0.0165	0.174
page_type	person	-1.776	0.7335	5.861	0.0155	0.169
open_q	1	0.204	0.0314	42.23	<.0001	1.226
multi_q	1	-0.1839	0.07916	5.398	0.0202	0.832
dropdown_q	1	0.7457	0.06147	147.1	<.0001	2.108
sec_q	1	3.189	0.05981	2843	<.0001	24.27

Table 6.11: The DQO-ordered questionnaires had fewer instances of breakoff than the fixed-order questionnaire.

Device type	λ =0.000	λ =0.005	λ =0.010	Fixed
Computer	150	144	144	176
Mobile	101	109	99	119

6.3.5 Limitations

Because the dataset used for these experiments was a convenience sample and not sampled to be representative of the entire population, we cannot claim that these results will generalize beyond this set of self-respondents on the online ACS questionnaire. For example, there are fewer Hispanic and nonwhite respondents in our dataset than in the U.S. population (7.6% of respondents in this sample are Hispanic, compared to 17.1% of the U.S. population, and 84.2% of respondents in this sample are white, compared to 76.1% of the U.S. population). The Internet self-respondents in this sample are also more educated than the general population: 27.0% of respondents have terminal bachelors degrees, and 19.8% have graduate degrees, compared to 19.0% bachelors degrees and 11.6% graduate degrees in the full population[5].

Furthermore, the dynamic ordering procedure was not tested on actual respondents and instead was simulated by reordering already collected data. The cognitive effects of reordering questions in a complex questionnaire like the ACS need to be explored. For example, the survival analysis models for breakoff in Section 6.3.3 show that respondents are highly likely to break off on pages with questions indicating a new section of the survey is beginning. Delaying these questions (or rewording them not to mention a new survey section) might delay breakoff, since respondents will not explicitly be told about the upcoming section of the survey. However, changing or delaying these questions could also cause respondents to be confused or frustrated if they cannot understand the structure of the questionnaire and their progress toward the end.

[5]Source: U.S. Census Bureau, 2015 American Community Survey 1-Year Estimates, accessed at `factfinder.census.gov`.

Table 6.12: Compared to the fixed-order questionnaire, DQO delayed breakoff by several questions.

Device type	λ =**0.000**	λ =**0.005**	λ =**0.010**
Computer	2.364	3.435	2.875
Mobile	0.849	1.171	1.453

6.3.6 Conclusion

In this application to the American Community Survey, we first modeled the relationship between respondent behavior (captured in paradata from the online survey instrument), page and question characteristics, and breakoff probability. Then we simulated adaptively ordering questions by using the DQO framework with utility defined as the probability that a respondent will answer a question on a page and cost defined as the time to answer a question. We found that, in this simulation, the DQO orderings had fewer instances of breakoff and more questions answered before breakoff than the fixed-order form that asked questions in the order of the standard ACS questionnaire.

6.4 Other potential survey applications

In this section, we elaborate on particular problems DQO could solve in other surveys. All surveys in this section are administered in computerized modes in which DQO would be possible. Some surveys, such as the Current Population Survey and the National Crime Victimization Survey, have predictions as goals (identifying employment status or classifying incidents of victimization) and can directly incorporate a prediction-motivated approach to DQO, as in Section 4.4.4. Other surveys, like the American Community Survey (ACS) and National Health Interview Survey, have the broad goal of collecting information on a population. DQO in these surveys would need to focus on maximizing information gain or respondent engagement, calculated from previously provided answers and paradata. National surveys are complicated, due to complex sampling requirements (*e.g.*, oversampling certain populations) and multi-purpose goals (*e.g.*, adding supplemental modules to core surveys).

Typically, statistical agencies already have strategies for handling nonresponse, and the current protocol could influence DQO. For example, the Census Bureau defines several categories of nonresponse in the ACS: households that answer basic demographic and housing questions, but not any detailed person questions, are considered "sufficient partial" responses who are not contacted for follow-up. Households that do not answer the basic questions are "insufficient partial" responses and are contacted for follow-up (U.S. Bureau of the Census, 2014a; Clark, 2014). Thus, Census might want to apply DQO within sections of the survey: first order questions in the basic subset, to ensure that more people will become at least sufficient partial responses, if not complete. Then, once a household's sufficient partial status is confirmed, the survey can continue with the detailed person questions.

6.4.1 Current Population Survey

The Current Population Survey (CPS) is a monthly survey of 60,000 households across the United States, jointly sponsored by the U.S. Census Bureau and Bureau of Labor Statistics (U.S. Bureau of the Census, 2006). Selected households are in the survey for four consecutive months, out of the survey for eight months, and then back in the survey for four months. The chief purpose of the CPS is to estimate the U.S. unemployment rate for the past month, and the majority of the official survey is devoted to this task. Respondents answer a battery of questions related to their work status in the past week to determine if they were employed, unemployed, or not in the labor force. There are over 200 questions in the labor force portion of the items. Not all of these questions apply to every household, so the current version of the CPS uses predefined skip patterns to avoid asking irrelevant questions. Augmenting the rule-based skip patterns with dynamic question ordering derived from statistical properties of the respondent could further lower respondent burden.

Various survey sponsors add supplemental question modules to the CPS (*e.g.*, the Tobacco Use Supplement, sponsored by the National Cancer Institute), which may also benefit from dynamic question ordering. The number of supplements is heavily restricted, due to not wanting to overburden respondents with too many questions and detract from the main purpose of the survey—estimating employment rate (U.S. Bureau of the Census, 2006). Using dynamic question ordering within or between modules could effectively select items to ask of populations of interest, thereby reducing the effective number of questions respondents must answer and increasing the potential for supplemental questions on the CPS.

6.4.2 National Health Interview Survey

The largest U.S. health survey, the National Health Interview Survey (NHIS) is administered in person to about 35,000 households each year (National Center for Health Statistics, 2014). The main purpose of the NHIS is to collect health information, at the household, family, and individual levels. Currently, the NHIS uses predefined skip patterns to advance respondents through the survey, but a statistical approach to question ordering could enhance the survey. Like the CPS, the NHIS has supplements to the main survey sponsored by other agencies. Supplemental questions are often asked in their own modules but are occasionally interspersed into the NHIS Core. As a CAPI-conducted survey, the NHIS could feasibly incorporate dynamic question ordering into its interview procedure.

The structure of the NHIS designates one person from a household as the "household respondent" who provides information for all members in the household (even for multi-family households). This type of proxy reporting is more likely to have errors than self-reports (Sudman et al., 1996), and so a dynamic question-ordering procedure would need to consider the impact of uncertain provided values when choosing which question to ask next.

The NHIS oversamples underrepresented populations, like black, Hispanic, and Asian people, to obtain more precise estimates for these populations (National Center for Health Statistics, 2014). With this goal in mind, DQO for the NHIS could consider the accuracy of imputed values for questions that are not yet asked, with thresholds for allowable imputation error. Such thresholds could be population-specific, with minorities' having lower allowable error thresholds, to ensure that more complete data are collected from these populations.

6.4.3 National Crime Victimization Survey

Every year the U.S. Census Bureau, on behalf of the Bureau of Justice Statistics, administers the National Crime Victimization Survey (NCVS) to 90,000 households (Bureau of Justice Statistics, 2014). Occupants of sampled households are interviewed every six months over three years. In interviews, respondents report victimizations in the previous six months. The interview is always conducted in a computerized mode (CAPI or CATI), with the first interview in person, so dynamic question ordering is possible in this computerized setting.

The NCVS collects detailed information about each incident reported by a respondent to classify incidents into fine-grained categories of crime (*e.g.*, "Robbery – attempted with injury"). The current NCVS design asks a respondent a set of questions regarding each incident they report and uses answers to these questions to classify the crime, rather than directly asking respondents for the crime category (Bureau of Justice Statistics, 2014). As such, this survey has a per-individual prediction problem at its core (labeling an incident as a type of crime), and could benefit from the prediction-guided DQO process set forth in Chapter 4.

DQO could further benefit the NCVS because, especially as households complete the survey multiple times, respondents recognize that reporting an incident results in an extended set of questions. This full questioning takes place for each individual report, including repeat victimizations (*e.g.*, domestic violence). To speed up the interview, participants are likely to underreport victimization, to avoid lengthy subsets of questions for each report. By reducing the number of questions to categorize incidents and using previously provided information to help in question ordering for repeat incidents, DQO could reduce the number of questions in the entire survey, making it less burdensome for respondents to provide complete reports.

6.5 Conclusion and future work

Dynamic question ordering—*i.e.*, choosing which question to ask a survey respondent next, depending on their answers to previous questions—can improve survey quality in two key ways. First, giving participants personalized question orders can engage them and motivate them to complete the survey. Second, eliciting the most relevant information for a particular respondent upfront can improve the quality of imputations for unanswered questions if the respondent breaks off before completing the questionnaire. For some surveys, the goal is only to estimate a value for each respondent. In this case, it is not even necessary for the participant to answer all questions—it is sufficient for them to answer a subset that will ensure a confident prediction.

This chapter adapted the DQO framework to two U.S. Census Bureau surveys and suggested ways to apply DQO to several other computerized national surveys. DQO sequentially considers which question to ask a respondent next, based on their previous answers, trading off the expected utility of having an answer to that question with the cost of asking that question. Thus, adaptive DQO to any particular survey requires defining both the utility and the cost of survey questions. The definition of "utility" for an answer depends on the survey and its purpose. Examples include information gain, response probability, (negative) breakoff probability, or certainty of the subsequent prediction. Similarly, the definition of question "cost" also depends on the survey. Examples include difficulty to the user of answering the question, (negative) likelihood of answering (since respondents may be reluctant to respond to sensitive questions, even if they are easy to answer), or breakoff rates of individual questions.

In this chapter on survey data collection, we illustrated the DQO framework on two surveys. First, we considered the Survey of Income and Program Participation (SIPP). We selected subsequent questions that maximized conditional entropy to get a "representative" set of answers from participants early on, trading off conditional entropy (utility) against item nonresponse rate (cost). We simulated survey breakoff and found that, compared to the standard SIPP question order, the dynamic order delays breakoff, better recovers values for unanswered items that must be imputed, and achieves better-quality survey estimates (lower bias and variance). Next, we considered the American Community Survey (ACS). The online mode of the ACS includes a rich set of paradata, tracking respondents' actions as they progress through the Internet questionnaire. We used these paradata to predict the likelihood of a respondent's breaking off on a certain page in the survey, given their previous behavior, and applied DQO, using respondent engagement (utility) and item response time (cost) to choose a question ordering for each respondent. Compared to the standard ACS question order, the dynamic order reduces and delays breakoff and therefore recovers more complete information from respondents. Finally, we discussed ways that dynamic question ordering could improve quality in computerized national surveys, focusing on unique aspects of each survey that DQO must take into account.

As more surveys move online or to computerized modes, dynamic question ordering can improve survey results at scale and at low cost to the data collectors. DQO trades off the utility from having an answer to a question with its cost and sequentially requests feature values in order to make useful, confident predictions and gather survey data with the resources users are willing and able to provide.

Although the supposed neutrality of the survey as an impartial data collection tool means that all respondents have the same (or very similar) survey experiences, this rigid structure can also hinder the natural flow of information that occurs in a conversation (Suchman & Jordan, 1990). Often for a participant, a particular event influences their answers for multiple questions. However, unless a direct question about this event appears in the survey, they have to answer many repetitive questions that could have been avoided in a conversation. Learning a latent structure of participants' answers in a survey could be a step toward uncovering these hidden events that determine the answers to multiple questions, and DQO could use this knowledge to guide question selection as well.

The cognitive aspects of survey methodology movement that originated in the 1980s (Jabine et al., 1984; Tanur, 1992) raised issues with the traditional approach to survey questionnaire design, which keeps order fixed for all respondents and which measures the same quantities at different points in time. The need to reduce respondent burden and to keep respondents engaged in online surveys is raising a complementary set of issues that are now being addressed under the rubric of adaptive survey design. These two perspectives do need to be reconciled in some fashion.

Given typical survey respondents' disengagement from surveys and declining survey response rates, maybe a new paradigm of survey collection, in which respondents get something useful to them out of answering a

survey, could motivate participants to provide complete and accurate responses (*e.g.*, (Marcus et al., 2007; Angelovska & Mavrikiou, 2013)). However, one clear downside to this approach is that giving respondents information that comes from the survey they are currently answering contaminates their response. For example, suppose that a person is answering questions about their energy-using habits to receive a personalized energy estimate, as in Section 4.4.4. Their current estimate for natural gas consumption is higher than they would like, and the next question asks for their preferred temperature in the winter. Because they do not want their estimate for natural gas usage to climb even higher, the respondent gives an optimistically low value for preferred temperature. The uncertainty associated with these predictions can also influence a user's decision to continue answering questions: once a participant feels that their given prediction is certain enough, they may stop answering questions. Depending on the purpose of the survey (namely, whether its chief goal is to provide information to or to collect information from the respondent), this type of breakoff may or may not be bad.

Chapter 7

Conclusion and future work

Thesis statement Dynamically ordering questions, based on already known information (such as previously answered questions or paradata), can lower the costs of data collection and prediction by tailoring the information-acquisition process to the individual. This thesis presents a framework for an iterative dynamic question ordering process that trades off question utility against question cost at data collection time. The exact metrics for utility and cost are application-dependent. We compare data quality and acquisition costs of our method to fixed-order approaches and show that our adaptive process obtains results of similar or better quality at lower cost.

7.1 Conclusion

This thesis presented a general framework for dynamically ordering questions, based on previously collected data, to engage respondents, improving prediction quality and survey completion. Our work considered two broad scenarios for data collection from users, systems, or survey respondents. In the first, we want to give the respondent a personalized prediction (Chapter 4) or score (Chapter 5), based on information they provide. Since it is possible to give a reasonable value with only a subset of questions, we are not concerned with motivating the user to provide values for all items. Instead, we want to order questions so that the user provides information that most improves the quality of our prediction or score, while not being too burdensome to answer. In the second scenario, our goal is to gather complete and accurate data (Chapter 6). Thus, we want to maximize survey completion (and the quality of necessary imputations). Therefore, we focus on ordering questions to engage the respondent and collect hopefully all the information we seek, or at least the information that most characterizes the respondent so that imputed values will be accurate.

In Chapter 4, we presented the FOCUS method (an instantiation of the DQO framework, using prediction certainty as the measure of item utility) for prediction-focused data collection. When gathering information to make a prediction, not all pieces of information may be relevant for a particular user at a particular time. Dynamically acquiring features for test-time prediction can reduce the costs of data collection while still maintaining the quality of the predictions, by adaptively selecting the most relevant features for a particular test sample. On all three validation datasets, FOCUS effectively lowered prediction costs (by reducing the number of additional, costly features to acquire), without sacrificing prediction quality for most values of n, the number of collected features, and λ, the utility-cost tradeoff parameter. The FOCUS framework's ability to support context-dependent costs (illustrated in the stress prediction example on the StudentLife dataset in Section 4.4.5) allows for richer, more realistic interpretations of feature cost, which may not be fixed for all test instances.

Chapter 5 considered a similarly output-focused data collection task, but for the application of scoring questionnaires to give a measurement of an underlying trait, such as intelligence or cultural values. We used the DQO framework to leverage previously provided information from a user to select which question is best to ask next, trading off the expected usefulness of that question (influence on the final score estimate) against its cost (response time). In simulated experiments on the pre-collected World Values Survey as well as a live deployment on the LabintheWild online experimental platform, we found that DQO achieves lower

error and lower cost than a fixed-order long form. Furthermore, DQO allows the questionnaire administrator to continue asking questions beyond a fixed short form's capacity, when a respondent is willing to spend more time or effort answering questions, resulting in improved score estimates relative to a short form.

Finally, Chapter 6 explored DQO in the survey-taking setting, where the goal is to collect complete and accurate information on a population. We illustrated how DQO can improve survey quality in two ways, on two different surveys. First, eliciting the most relevant information for a particular respondent upfront can improve the quality of imputations for unanswered questions if the respondent breaks off before completing the questionnaire. We illustrated this benefit on the Survey of Income and Program Participation (SIPP) by trading off the maximum conditional entropy of a new question (utility) against that question's nonresponse rate (cost). We simulated survey breakoff and found that, compared to the standard SIPP question order, the dynamic order delays breakoff, better recovers values for unanswered items that must be imputed, and achieves better quality survey estimates (lower bias and variance). Second, giving participants personalized question orders can engage them and motivate them to complete the survey. We illustrated this benefit on the Internet questionnaire for the American Community Survey (ACS). We used ACS paradata to predict the likelihood of a respondent's breaking off on a certain page in the survey, given their previous behavior, and applied DQO, using respondent engagement (utility) and item response time (cost) to choose a question ordering for each respondent. Compared to the standard ACS question order, the dynamic order delays breakoff and recovers more complete information from respondents.

7.2 Future work

One key area for future research in adaptive, cost-effective data collection is to combine multiple goals in data collection. The work presented in this thesis used a single objective in ordering questions for a respondent—improving the quality of a prediction, increasing the information content in a set of responses, or engaging a respondent—but many real-world data collection problems have multiple simultaneous goals. For example, an intelligent personal assistant wants to manage a user's schedule while monitoring their activity and also making recommendations and needs to collect data from various sources (sensors, the Internet, the user, *etc.*) to meet all these goals. Large-scale government surveys are often co-sponsored by multiple federal agencies who want information from different survey modules (*e.g.*, questions about health status versus questions about income). In both of these examples, data need to be collected in a way that adequately meets all subgoals without exhausting a shared budget.

Applying dynamic data collection to other domains will also further research in this area. The generalizability of this dynamic item ordering framework means that it often cannot be directly applied to a new problem without domain-specific adjustments. Different communities care about different aspects of data collection, and addressing those concerns is necessary for realistic solutions in those communities. These practical considerations highlight facets of the framework that also apply to other situations. For example, the concept of context-dependent costs is relevant to mobile sensing applications where the battery context of a sensor (*e.g.*, smartphone) influences the practical cost consideration of different features to gather can also apply to the concern about order effects in survey methodology. When people interpret questions differently in different orders, survey researchers often recommend standard orders for certain questions (Sudman et al., 1996). This constrained order can be enforced in a dynamic ordering by considering the cost of a dependent question to be infinite in the context of its prerequisite question(s) not being asked yet.

Another area for future work is in scaling up data collection. The applications in this thesis considered all potential items that might be acquired next when calculating the utility-cost objective in Equation 3.1. When there are very many items that can be collected, performing this calculation for every remaining item at each step in data collection can be time-consuming. In this case, being strategic about covering the search space can reduce the computation time to a reasonable amount. For example, if the set of items is structured in a way that some items will always be worse than others, then the calculations can be made over a reduced set of items. Alternatively, reusing or making inexpensive updates to previous calculations could also reduce the computation time for selecting a new item.

The work in this thesis on dynamic question ordering is just one aspect of the larger field of cost-effective data collection. The results of the DQO framework for dynamic data collection in both prediction and survey collection illustrate the cost-saving benefit of tailoring data collection to an individual data point and

considering the usefulness of different pieces of information for the task at hand when choosing which to acquire next. This work in cost-effective test-time data collection is also related to paradigms of training-time data collection, such as active learning and online learning. The applications presented in this thesis all relied on complete training data to understand the data (*e.g.*, learn a predictive model, distributions of responses, breakoff probabilities). Relaxing this requirement and making use of adaptive data collection methods from the start (without needing complete training data) will further the cost savings in all stages of data collection.

There are always limits to data collection, whether they are of time, money, battery, bandwidth, privacy, *etc.* Strategically working within these limits, using cost-effective data collection, ensures that useful data can still be gathered.

References

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98.

Angelovska, J., & Mavrikiou, P. M. (2013). Can creative web survey questionnaire design improve the response quality? *University of Amsterdam, AIAS Working Paper*, *131*.

Aronson, Z. H., Reilly, R. R., & Lynn, G. S. (2006). The impact of leader personality on new product development teamwork and performance: The moderating role of uncertainty. *Journal of Engineering and Technology Management*, *23*(3), 221–247.

Ballesteros, M., & Bohnet, B. (2014). Automatic feature selection for agenda-based dependency parsing. In *COLING* (pp. 794–805).

Banerjee, N., Rahmati, A., Corner, M. D., Rollins, S., & Zhong, L. (2007). Users and batteries: Interactions and adaptive energy management in mobile systems. In *International conference on ubiquitous computing* (pp. 217–234).

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*(2), 182–200.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493–513.

Bartholomew, D. J., Steele, F., Galbraith, J., & Moustaki, I. (2008). *Analysis of multivariate social science data*. CRC press.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*(3), 390–399.

Beaumont, J.-F., Bocci, C., & Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, *30*(4), 607–621.

Beigl, M. (1999). Point & click – interaction in smart environments. In *International symposium on handheld and ubiquitous computing* (pp. 311–313).

Benedetto, G., Stinson, M., & Abowd, J. M. (2013). *The creation and use of the SIPP synthetic beta.* www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf. U.S. Census Bureau.

Bernard, L. C., Walsh, R. P., & Mills, M. (2005). Ask once, may tell: Comparative validity of single and multiple item measurement of the Big-Five personality factors. *Counseling and Clinical Psychology Journal*, *2*(1), 40–57.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bradburn, N. (1978). Respondent burden. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 35–40).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Budde, M., Berning, M., Baumgärtner, C., Kinn, F., Kopf, T., Ochs, S., ... Beigl, M. (2013). Point & control–interaction in smart environments: You only click twice. In *Proceedings of the 2013 ACM conference on pervasive and ubiquitous computing adjunct publication* (pp. 303–306).

Bureau of Justice Statistics. (2014). *National Crime and Victimization Survey: Technical documentation.* Retrieved 2017-04-14, from `http://www.bjs.gov/content/pub/pdf/ncvstd13.pdf`

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, *226*(1), 115–121.

Chang, H.-H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*(4), 333–341.

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in α-stratified computerized adaptive testing. *Applied Psychological Measurement*, *27*(4), 262–274.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229.

Chang, H.-H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? balancing influence of initial items in adaptive testing. *Psychometrika*, *73*(3), 441–450.

Chang, T.-H., & Li, Y. (2011). Deep shot: A framework for migrating tasks across devices using mobile phone cameras. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2163–2172).

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (1999). Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing.

Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, *69*(1), 35–49.

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467–482.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62–79.

Clark, S. L. (2014). American Community Survey item nonresponse rates: Mail versus internet. *American Community Survey Research and Evaluation Program (March)*. Retrieved 2017-04-14, from `https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Clark_01.pdf`

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2010). The impact of progress indicators on task completion. *Interacting with computers*, *22*(5), 417–427.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. doi: 10.1177/00131640021970934

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory: Manual, form S and form R.* Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1989). The NEO-PI/NEO-FFI manual supplement. *Odessa, FL: Psychological Assessment Resources*, *40*.

Couper, M., & Wagner, J. R. (2012). Using paradata and responsive design to manage survey nonresponse. In *58th world statistical congress* (pp. 542–548).

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the survey research methods section of the American Statistical Association* (pp. 41–49).

Couper, M. P., Alexander, G. L., Zhang, N., Little, R. J., Maddy, N., Nowak, M. A., ... Johnson, C. C. (2010). Engagement and retention: Measuring breadth and depth of participant use of an online intervention. *Journal of Medical Internet Research*, *12*(4), e52.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*, 187–220.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001). Web surveys: Perceptions of burden. *Social Science Computer Review*, *19*(2), 146–162.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874.

Dahl, R. A. (1997). Development and democratic culture. *Consolidating the Third Wave Democracies*, *34*.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

De Ayala, R. J. (2013). *The theory and practice of item response theory.* Guilford Publications.

de Freitas, A., Nebeling, M., Chen, X. A., Yang, J., Ranithangam, A. S. K. K., & Dey, A. K. (2016). Snap-to-it: A user-inspired platform for opportunistic device interactions. In *Proceedings of the 34th annual ACM conference on human factors in computing systems (chi 2016).*

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage Publications.

Dietz, T., Gardner, G. T., Gilligan, J., Stern, P. C., & Vandenbergh, M. P. (2009). Household actions can provide a behavioral wedge to rapidly reduce U.S. carbon emissions. In *Proceedings of the National Academy of Sciences* (Vol. 106, pp. 18452–18456). National Academy of Sciences.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203.

Douthitt, R. A. (1989). An economic analysis of the demand for residential space heating fuel in Canada. *Energy*, *14*(4), 187–197.

Early, K., Fienberg, S. E., & Mankoff, J. (2016). Test-time feature ordering with FOCUS: Interactive predictions with minimal user burden. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 992–1003.

Early, K., Mankoff, J., & Fienberg, S. E. (2017). Dynamic question ordering in online surveys. *Journal of Official Statistics (to appear)*.

Ferreira, D., Ferreira, E., Goncalves, J., Kostakos, V., & Dey, A. K. (2013). Revisiting human-battery interaction with an interactive battery interface. In *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing* (pp. 563–572).

Ferreira, D., Kostakos, V., & Dey, A. K. (2015). AWARE: Mobile context instrumentation framework. *Frontiers in ICT*, *2*, 6.

Ferreira, P., Sanches, P., Höök, K., & Jaensson, T. (2008). License to chill!: How to empower users to cope with stress. In *Proceedings of the 5th Nordic conference on human-computer interaction: building bridges* (pp. 123–132).

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. T. (2013). Watson: Beyond jeopardy! *Artificial Intelligence*, *199*, 93–105.

Fox, R. J., Crask, M. R., & Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, *52*(4), 467–491.

Fricker, S., Yan, T., & Tsai, S. (2014). Response burden: What predicts it and who is burdened out. In *JSM Proceedings* (pp. 4568–4577).

Friedman, J. H., Kohavi, R., & Yun, Y. (1996). Lazy decision trees. In *AAAI/IAAI, vol. 1* (pp. 717–724).

Fuller, W. A. (2009). *Measurement error models*. New York: Wiley.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313–328.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. doi: 10.1093/poq/nfp031

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. J. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, *12*(1), 83–104.

Golbandi, N., Koren, Y., & Lempel, R. (2011). Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 595–604).

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*(1), 26.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.

Gonzalez, J. M., & Eltinge, J. L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. In *JSM proceedings.*

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Goyder, J., Warriner, K., & Miller, S. (2002). Evaluating socio-economic status (SES) bias in survey nonresponse. *Journal of Official Statistics*, *18*(1), 1.

Gray, R., Campanelli, P., Deepchand, K., & Prescott-Clarke, P. (1996). Exploring survey non-response: The effect of attrition on a follow-up of the 1984-85 Health and Life Style Survey. *The Statistician*, 163–183.

Griffin, D., & Nelson, D. (2014). *Reducing respondent burden in the ACS's computer assisted personal visit interviewing operation – phase 1 results.* https://www.census.gov/content/dam/Census/library/working-

papers/2014/acs/2014_Griffin_02.pdf.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

Groves, R. M., & Heeringa, S. G. (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*(3), 439–457.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*(1), 2–31.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *The Public Opinion Quarterly*, *64*(3), 299–308.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hallak, A., Di Castro, D., & Mannor, S. (2015). Contextual Markov decision processes. *arXiv preprint arXiv:1502.02259*.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer.

Hauser, R. M. (2005). Survey response in the long run: The Wisconsin Longitudinal Study. *Field Methods*, *17*(1), 3–29.

He, H., Daumé III, H., & Eisner, J. (2012). Cost-sensitive dynamic feature selection. In *Inferning 2012: ICML workshop on interaction between inference and learning*.

He, H., Daumé III, H., & Eisner, J. (2013). Dynamic feature selection for dependency parsing. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1455–1464).

Healey, B. (2007). Drop downs and scroll mice: The effect of response option format and input mechanism employed on data quality in web surveys. *Social Science Computer Review*, *25*(1), 111–128.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*(3), 360–373.

Hirschberg, P. A., Abrams, E., Bleistein, A., Bua, W., Delle Monache, L., Dulong, T. W., ... others (2011). A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bulletin of the American Meteorological Society*, *92*(12), 1651.

Horwitz, R., Tancreto, J., Zelenak, M. F., & Davis, M. (2012). *Data quality assessment of the American Community Survey Internet response data*.

Hudd, S. S., Dumlao, J., Erdmann-Sager, D., Murray, D., Phan, E., Soukas, N., & Yokozuka, N. (2000). Stress at college: Effects on health habits, health status and self-esteem. *College Student Journal*, *34*(2), 217–228.

Inglehart, R., & Welzel, C. (2010). Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, *8*(02), 551–567.

Jabine, T. B., Straf, M. L., Tanur, J. M., & Tourangeau, R. (Eds.). (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academies Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five inventory – versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Johnson, T. P., O'Rourke, D., Burris, J., & Owens, L. (2002). Culture and survey nonresponse. *Survey Nonresponse*, 55–69.

Kaczmirek, L. (2008). *Human-survey interaction: Usability and nonresponse in online surveys* (Doctoral dissertation, Universität Mannheim). http://ub-madoc.bib.uni-mannheim.de/2150/1/kaczmirek2008.pdf.

Kakade, S. M. (2003). *On the sample complexity of reinforcement learning* (Unpublished doctoral dissertation).

Kaldenberg, D. O., Koenig, H. F., & Becker, B. W. (1994). Mail survey response rate patterns in a population of the elderly: Does response deteriorate with age? *The Public Opinion Quarterly*, *58*(1), 68–76.

Kalia, M. (2002). Assessing the economic impact of stress – the modern day hidden epidemic. *Metabolism*, *51*(6), 49–53.

Kamakura, W. A., & Balasubramanian, S. K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment*, *53*(3), 502–519. doi: 10.1207/s15327752jpa5303_8

Kapelner, A., & Chandler, D. (2010). Preventing satisficing in online surveys. In *Proceedings of CrowdConf*.

Kaza, N. (2010). Understanding the spectrum of residential energy consumption: A quantile regression approach. *Energy Policy*, *38*(11), 6574–6585.

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 759–779.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: Techniques for censored and truncated data*. Springer Science & Business Media.

Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: Does completion device affect survey responses? *Research in Higher Education*, *56*(2), 166–177.

Laurie, H., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, *15*(2), 269–282.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000). *Content balancing in stratified computerized adaptive testing designs*. ERIC Clearinghouse.

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review*, *53*(01), 69–105.

Liu, Q., McEvoy, P., Kimber, D., Chiu, P., & Zhou, H. (2006). On redirecting documents with a mobile camera. In *2006 IEEE workshop on multimedia signal processing* (pp. 467–470).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 1150–1157).

Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., & Campbell, A. T. (2010). The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM conference on embedded networked sensor systems* (pp. 71–84).

Lundquist, P., & Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish living conditions survey. *Journal of Official Statistics*, *29*(4), 557–582.

Manfreda, K. L., & Vehovar, V. (2002). Survey design features influencing response rates in web surveys. In *The International Conference on Improving Surveys*. Copenhagen, Denmark. Retrieved 2017-04-14, from `http://www.websm.org/uploadi/editor/Lozar_Vehovar_2001_Survey_design.pdf`

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys a field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383. doi: 10.1177/0894439307297606

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. *New horizons in testing*, 223–226.

McCutcheon, A. L., al Baghal, T., & Tsabutashvili, D. (2013). Predicting survey breakoff in internet survey panels. In *AAPOR*. Boston, Massachusetts.

McFarland, S. G. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*(2), 208–215.

Min, C., Yoo, C., Hwang, I., Kang, S., Lee, Y., Lee, S., ... Song, J. (2015). Sandra helps you learn: The more you walk, the more battery your phone drains. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 421–432).

Misra, R., & McKean, M. (2000). College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies*, *16*(1), 41.

Moberg, G. P. (2000). Biological response to stress: Implications for animal welfare. *The biology of animal stress: Basic principles and implications for animal welfare*, 1–21.

Montgomery, J. M., & Cutler, J. (2013). Computerized adaptive testing for public opinion surveys. *Political Analysis*, *21*(2), 172–192.

National Center for Health Statistics. (2014). *Survey description: National Health Interview Survey*. Retrieved 2017-04-14, from `http://ftp.cdc.gov/pub/Health_Statistics/NCHs/Dataset_Documentation/NHIS/2014/srvydesc.pdf`

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*(350), 351–356.

Pattuk, E., Kantarcioglu, M., Ulusoy, H., & Malin, B. (2015). Privacy-aware dynamic feature selection. In *2015 IEEE 31st international conference on data engineering* (pp. 78–88).

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97.

Peytchev, A. (2011). Breakoff and unit nonresponse across web surveys. *Journal of Official Statistics*, *27*(1), 33.

Porter, S. R. (2004). Raising response rates: What works? *New Directions for Institutional Research*, *2004*(121), 5–21.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Ravi, N., Scott, J., Han, L., & Iftode, L. (2008). Context-aware battery management for mobile phones. In *Proceedings of the sixth annual IEEE international conference on pervasive computing and communications* (pp. 224–233).

Reinecke, K., & Gajos, K. Z. (2015). LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 1364–1378).

World Values Survey Association. (2015). *World Values Survey Wave 6 2010 – 2014 official data file v.20150418*. Aggregate File Producer: ASEP/JDS, Madrid SPAIN. `www.worldvaluessurvey.org`.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327.

Ross, S. E., Niebling, B. C., & Heckert, T. M. (1999). Sources of stress among college students. *Social psychology*, *61*(5), 841–846.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, *8*(Jul), 1623–1657.

Samadi, M., Talukdar, P., Veloso, M., & Mitchell, T. (2015). Askworld: Budget-sensitive query evaluation for knowledge-on-demand. In *Proceedings of the 24th international conference on artificial intelligence* (pp. 837–843).

Särndal, C.-E., & Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, *2*(4), 361–387.

Särndal, C.-E., & Lundquist, P. (2014). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société Française de Statistique*, *155*(4), 28–50.

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*(3), 506–516.

Sauermann, H., & Roach, M. (2013). Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, *42*(1), 273–286.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, *39*(1), 29–58.

Seryak, J., & Kissock, K. (2003). Occupancy and behavioral effects on residential energy use. In *Proceedings of the Solar Conference* (pp. 717–722).

Sharma, N., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, *108*(3), 1287–1301.

Shi, T., Steinhardt, J., & Liang, P. (2015). Learning where to sample in structured prediction. In *AISTATS*.

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20*(3), 249–271.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167–194.

Strubell, E., Vilnis, L., Silverstein, K., & McCallum, A. (2015). Learning dynamic feature selection for fast sequential prediction. *arXiv preprint arXiv:1505.06169*.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews (with discussion). *Journal of the American Statistical Association*, *85*(409), 232–253.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Suh, H., Shahriaree, N., Hekler, E. B., & Kientz, J. A. (2016). Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 3988–3999).

Sun, M., Li, F., Lee, J., Zhou, K., Lebanon, G., & Zha, H. (2013). Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the sixth ACM international conference on web search and data mining* (pp. 445–454).

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*(8), 1819–1835.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the military testing association* (pp. 973–977).

Tanur, J. M. (1992). *Questions about questions: Inquiries into the cognitive bases of surveys.* New York: Russell Sage Foundation.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academies Press.

Tourangeau, R., Michael Brick, J., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(1), 203–223.

Trapeznikov, K., & Saligrama, V. (2013). Supervised sequential classification under budget constraints. In *AISTATS* (pp. 581–589).

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, *50*(10), 2231–2242.

U.S. Bureau of the Census. (2006). *Design and methodology: Current Population Survey.*

U.S. Bureau of the Census. (2014a). *American Community Survey: Design and methodology.*

U.S. Bureau of the Census. (2014b). *Survey of Income and Program Participation.* Retrieved 2016-02-05, from `www.census.gov/programs-surveys/sipp`

U.S. Bureau of the Census. (2016). *American Community Survey.* Retrieved 2016-02-22, from `http://www2.census.gov/programs-surveys/acs/methodology/questionnaires/2016/quest16.pdf`

U.S. Energy Information Administration. (2009). *Residential Energy Consumption Survey 2009.* `www.eia.gov/consumption/residential/data/2009/`.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing.* Springer.

Van Goor, H., & Rispens, S. (2004). A middle class image of society. *Quality and Quantity*, *38*(1), 35–49.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203–226.

Wagner, J. R. (2008). *Adaptive survey design to reduce nonresponse bias* (Unpublished doctoral dissertation). University of Michigan.

Wagner, J. R. (2012). Adaptive contact strategies in telephone and face-to-face surveys. In *Survey research methods* (Vol. 7, pp. 45–55).

Wagner, J. R., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, *28*(4), 477.

Wagner-Menghin, M. M. (2002). Towards the identification of non-scalable personality questionnaire respondents: Taking response time into account. *Psychological Test and Assessment Modeling*, *44*(1), 62.

Walston, J. T., Lissitz, R. W., & Rudner, L. M. (2006). The influence of web-based questionnaire presentation variations on survey cooperation and perceptions of survey quality. *Journal of Official Statistics*, *22*(2), 271–291.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., ... Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3–14).

Weisberg, S. (2014). *Applied linear regression* (4th ed.). New York: Wiley.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375.

Weiss, D. J., & Taskar, B. (2013). Learning adaptive value of information for structured prediction. In *Advances in neural information processing systems* (pp. 953–961).

Welzel, C. (2013). *Freedom rising: Human empowerment and the quest for emancipation.* Cambridge University Press.

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, *19*(5), 373–390.

Wu, J., Pan, G., Zhang, D., Li, S., & Wu, Z. (2010). MagicPhone: Pointing & interacting. In *Proceedings of the 2010 ACM conference on pervasive and ubiquitous computing adjunct publication* (pp. 451–452).

Xu, Z. E., Kusner, M. J., Weinberger, K. Q., Chen, M., & Chapelle, O. (2014). Classifier cascades and trees for minimizing feature evaluation cost. *Journal of Machine Learning Research*, *15*(1), 2113–2144.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68.

Yu, E. C., Fricker, S., & Kopp, B. (2015). Can survey instructions relieve respondent burden? In *AAPOR.*

Appendix A

Algorithms

Algorithm 1 Estimating values z for still-unknown features

Input: $X \in \mathbb{R}^{n \times d}, x \in \mathbb{R}^d, \mathcal{K} \subseteq \{1, ..., d\}, k \in \mathbb{Z}^+$
Output: $z \in \mathbb{R}^d$

function ESTIMATE_FEATURES($X, x, \mathcal{K}, k$)
    $z_{\mathcal{K}} \leftarrow x_{\mathcal{K}}$ ▷ Copy over the known features
    $\mathcal{I} \leftarrow$ get_knn($X_{:,\mathcal{K}}, z_{\mathcal{K}}, k$) ▷ Index $z_{\mathcal{K}}$'s kNNs
    for $f \in \{1, ..., d\} \setminus \mathcal{K}$ **do** ▷ For all unknown f
        $z_f \leftarrow \text{mean}(X_{\mathcal{I},f})$ ▷ Estimate z_f from kNNs' values for feature f
    end for
    return z
end function

Algorithm 2 Calculating the expected prediction uncertainty for each candidate feature

Input: $\mathcal{K} \subseteq \{1, ..., d\}$, $z \in \mathbb{R}^d$, feat_ranges, feat_proportions, model
Output: $E \in \mathbb{R}^d$

function EXPECTED_UNCERTAINTY($\mathcal{K}$, z, feat_ranges, feat_proportions, model)
    $E \leftarrow \mathbf{0}^d$
    for $f \in \{1, ..., d\} \setminus \mathcal{K}$ **do** ▷ For all unknown f
        $R \leftarrow$ feat_ranges$\{f\}$, $u \leftarrow \mathbf{0}^{|R|}$
        for $\ell \in \{1, ..., |R|\}$ **do** ▷ For each value f can take on
            $\bar{z} \leftarrow z$, $\bar{z}_f \leftarrow R_\ell$ ▷ f-th feature is assigned
            $u_\ell \leftarrow$ CALCULATE_UNCERTAINTY($\bar{z}$, model)
        end for
        $p \leftarrow$ feat_proportions$\{f\}$
        $E_f \leftarrow p^T u$ ▷ Expectation of prediction uncertainty
    end for
    return E
end function

Algorithm 3 Dynamically choosing a question ordering $\mathcal{A}$ and making a sequence of predictions $\hat{y}$ at the current feature values and estimates as feature values are provided

Input: $X \in \mathbb{R}^{n \times d}$, $x \in \mathbb{R}^d$, $\mathcal{K} \subseteq \{1, ..., d\}$, feat_ranges, feat_proportions, $\lambda \in \mathbb{R}$, $c \in \mathbb{R}^d$, model
Output: $\mathcal{A} \subseteq \{1, ..., d\}, \hat{y} \in \mathbb{R}^{|\mathcal{A}|+1}$

function DQO_ALL(X, x, $\mathcal{K}$, k, δ, α, feat_ranges, feat_proportions, $\hat{\beta}$, $\hat{\sigma}^2$, λ, c)
  $\mathcal{A} \leftarrow \{\}$, $\hat{y} \leftarrow \{\}$
  for $i \in \{1, ..., d - |\mathcal{K}|\}$ **do**
    $z \leftarrow$ ESTIMATE_FEATURES$(X, x, \mathcal{K})$
    $\hat{y}_i \leftarrow$ PREDICT(z, model) ▷ Predict on features and estimates
    $E \leftarrow$ EXPECTED_UNCERTAINTY(
      X, $\mathcal{K}$, z, feat_ranges, feat_proportions, model)
    $f^\star \leftarrow \arg\min_{f \notin \mathcal{K}} (E_f + \lambda \cdot c_f)$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{f^\star\}, \mathcal{K} \leftarrow \mathcal{K} \cup \{f^\star\}$
    $z_{f^\star} \leftarrow x_{f^\star}$ ▷ Ask and receive value for $f^\star$
  end for
  $z \leftarrow$ ESTIMATE_FEATURES$(X, x, \mathcal{K})$
  $\hat{y}_{d-|\mathcal{K}|+1} \leftarrow$ PREDICT(z, model) ▷ Make final prediction
  return $\mathcal{A}, \hat{y}$
end function

Appendix B

Experiment details

B.1 Residential Energy Consumption Survey

The results presented in this thesis come from the 2009 Residential Energy Consumption Survey (RECS) microdata, which are available online[1]. The 2015 RECS dataset was recently released and is also available online[2].

The cost categories assigned to all the features in RECS are presented in Table B.1. Explanations for the meanings of the cost categories are given in Table 4.2.

Table B.1: Feature cost categories for all RECS variables.

Cost	Feature
0	Census Region
0	Census Division
0	Reportable states and groups of states
0	Type of housing unit
0	Heating degree days in 2009, base temperature 65F
0	Cooling degree days in 2009, base temperature 65F
0	Heating degree days, 30-year average 1981-2010, base 65F
0	Cooling degree days, 30-year average 1981-2010, base 65F
0	Building America Climate Region
0	AIA Climate Zone, based on average temperatures from 1981 - 2010
0	Housing unit in Census Metropolitan Statistical Area or Micropolitan Statistical Area
0	Housing unit classified as urban or rural by Census
0	Housing unit is owned, rented, or occupied without payment of rent
3	Housing unit part of condominium or cooperative
4	Year housing unit was built
4	Year range when housing unit was built
1	Year range when household moved in
3	Converted 2-4 unit apartment building
5	Converted 2-4 unit apartment building was originally a single-family house
5	Converted 2-4 unit apartment building more like single family house or apartment building
3	Number of floors in a 5+ unit apartment building
3	Number of apartment units in a 5+ unit apartment building
5	Major outside wall material
6	Major roofing material

Continued on next page

[1] https://www.eia.gov/consumption/residential/data/2009/
[2] https://www.eia.gov/consumption/residential/data/2015/

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
0	Studio apartment
3	Number of floors in an apartment
3	Number of stories in a single-family home
3	Addition to a mobile home
0	Number of bedrooms
0	Number of full bathrooms
3	Number of half bathrooms
3	Number of rooms other than bedroom(s) and bathroom(s)
3	Total number of rooms in the housing unit
3	Basement in housing unit
6	Housing unit over a crawl space
6	Housing unit over a concrete slab
3	Finished basement
3	Number of finished rooms in the basement
3	Heating used in basement
3	All or partial basement heating
6	Portion of the basement which is heated
3	Cooling used in basement
3	All or partial basement cooling
6	Portion of the basement which is cooled
3	Portion of basement exclusively used by housing unit in apartment building with 2-4 units
3	Attic in housing unit
3	Finished attic
3	Number of finished rooms in the attic
3	Heating used in attic
3	All or partial attic heating
6	Portion of the attic which is heated
3	Cooling used in attic
3	All or partial attic cooling
6	Portion of the attic which is cooled
3	Portion of attic exclusively used by housing unit in apartment building with 2-4 units
3	Attached garage
3	Size of attached garage
3	Location of attached garage
3	Heating used in attached garage
3	Cooling used in attached garage
3	Detached garage or carport
3	Size of detached garage or carport
5	Outlet within 20 feet of vehicle parking
3	Number of stoves (one appliance with cooktop and an oven)
5	Fuel used by most-used stove
5	Number of separate cooktops
5	Fuel used by most-used separate cooktop
5	Number of separate ovens
5	Fuel used by separate oven
1	Frequency of oven use
5	Self-cleaning oven
5	Continuous or manual cleaning cycle for most-used oven
1	Microwave oven used
1	Microwave used for defrosting
5	Outdoor grill used

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
5	Fuel used by outdoor grill
5	Built-in indoor grill used
5	Fuel used by built-in indoor grill
1	Toaster used
1	Frequency hot meals are cooked
1	Most-used cooking fuel
1	Coffee maker used
3	Number of refrigerators used
3	Door arrangement of most-used refrigerator
3	Size of most-used refrigerator
3	Defrosting type of most-used refrigerator
3	Through-the-door ice and water on most-used refrigerator
7	Age of most-used refrigerator
7	Energy Star most-used refrigerator
7	Most-used refrigerator replaced by this household in the last 4 years
7	Assistance for replacing most-used refrigerator
7	Year of assistance for most-used refrigerator
3	Door arrangement of second most-used refrigerator
3	Size of second most-used refrigerator
3	Defrosting type of second most-used refrigerator
7	Number of months second most-used refrigerator used in 2009
7	Age of second most-used refrigerator
7	Energy Star second most-used refrigerator
3	Door arrangements of third most-used refrigerator
3	Size of third most-used refrigerator
3	Defrosting type of third most-used refrigerator
7	Number of months third most-used refrigerator used in 2009
7	Age of third most-used refrigerator
7	Energy Star third most-used refrigerator
3	Separate freezer used
3	Number of separate freezers used
5	Type of most-used freezer
5	Size of most-used freezer
3	Defrosting type for most-used freezer
7	Age of most-used freezer
7	Most-used freezer replaced by this household in the last 4 years
7	Assistance for replacing most-used freezer
7	Year of assistance for most-used freezer
5	Type of second most-used freezer
5	Size of second most-used freezer
3	Defrosting type for second most-used freezer
7	Age of second most-used freezer
3	Dishwasher used
1	Frequency dishwasher used
7	Age of dishwasher
7	Energy Star dishwasher
7	Dishwasher replaced by this household in the last 4 years
7	Assistance for replacing dishwasher
7	Year of assistance for dishwasher
3	Clothes washer used in home
5	Top or front loading clothes washer used in home

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
1	Frequency clothes washer used
1	Water temperature used for clothes washer wash cycle
1	Water temperature used for clothes washer rinse cycle
7	Age of clothes washer
7	Energy Star clothes washer
7	Clothes washer replaced by this household in the last 4 years
7	Assistance for replacing clothes washer
7	Year of assistance for the clothes washer
3	Clothes dryer used in home
5	Fuel used by clothes dryer
1	Frequency clothes dryer used
7	Age of clothes dryer
1	Number of televisions used
2	Size of most-used TV
2	Display type of most-used TV
2	Cable box or satellite box connected to the most-used TV
2	DVR built into the cable box or satellite box connected to the most-used TV
2	Separate DVR connected to the most-used TV
2	Digital converter box connected to the most-used TV
2	Video game console connected to the most-used TV
2	Combo VCR/DVD connected to the most-used TV
2	VCR connected to the most-used TV
2	DVD player connected to the most-used TV
2	Home theater system connected to the most-used TV
2	Other set-top box connected to the most-used TV
2	Size of second most-used TV
2	Display type of second most-used TV
2	Cable box or satellite box connected to the second most-used TV
2	DVR built into the cable box or satellite box connected to the second most-used TV
2	Separate DVR connected to the second most-used TV
2	Digital converter box connected to the second most-used TV
2	Video game console connected to the second most-used TV
2	Combo VCR/DVD connected to the second most-used TV
2	VCR connected to the second most-used TV
2	DVD player connected to the second most-used TV
2	Home theater system connected to the second most-used TV
2	Other set-top box connected to the second most-used TV
2	Size of third most-used TV size
2	Display type of third most-used TV
2	Cable box or satellite box connected to the third most-used TV
2	DVR built into the cable box or satellite box connected to the third most-used TV
2	Separate DVR connected to the third most-used TV
2	Digital converter box connected to the third most-used TV
2	Video game console connected to the third most-used TV
2	Combo VCR/DVD connected to the third most-used TV
2	VCR connected to the third most-used TV
2	DVD player connected to the third most-used TV
2	Home theater system connected to the third most-used TV
2	Other set-top box connected to the third most-used TV
1	Computer used at home
1	Number of computers used

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
1	Most-used computer - desktop or laptop
1	Monitor type of most-used computer
1	Turn off most-used computer when not in use
1	Sleep or standby mode for most-used computer when not in use
1	Second most-used computer - desktop or laptop
1	Monitor type of second most-used computer
1	Turn off second most-used computer when not in use
1	Sleep or standby mode for second most-used computer when not in use
1	Third most-used computer - desktop or laptop
1	Monitor type of third most-used computer
1	Turn off third most-used computer when not in use
1	Sleep or standby mode for third most-used computer when not in use
1	Internet access at home
1	Dial-up internet access
1	DSL or Fiber Optic internet access
1	Cable internet access
1	Satellite internet access
1	Wireless internet in home
1	Number of printers used
1	Separate fax machine used
1	Separate copy machine used
6	Well water pump used
6	Automotive block or engine heater or battery blanket used
6	Evaporative cooler used
1	Large heated aquarium used
1	Stereo equipment used
1	Cordless telephone used
1	Answering machine used
2	Number of rechargeable tools and appliances used
1	Charging patterns for rechargeable tools and appliances
1	Chargers for rechargeable tools and appliances left plugged into wall
2	Number of rechargeable electronic devices used
1	Charging patterns for rechargeable electronic devices
1	Chargers for rechargeable electronic devices left plugged into wall
5	Space heating equipment used
5	No space heating equipment, or unused space heating equipment
5	Unused space heating equipment type
5	Fuel for unused space heating equipment
3	Type of main space heating equpiment used
3	Main space heating fuel
3	Routine service or maintenance performed on main space heating equipment
7	Age of main space heating equipment
7	Main space heating equipment replaced by this household in the last 4 years
7	Assistance for replacing or maintaining main space heating equipment
7	Year of assistance for main space heating equipment
3	Main space heating equipment heats other homes, business, or farm
5	Secondary space heating equipment used
5	Heat pump used for secondary space heating
5	Central warm-air furnace used for secondary space heating
5	Fuel used by warm-air furnace for secondary space heating
5	Hot water system used for secondary space heating

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
5	Fuel used by hot water system for secondary space heating
5	Built-in electric units used for secondary space heating
5	Pipeless furnace used for secondary space heating
5	Fuel used by pipeless furnace for secondary space heating
5	Built-in room heaters used for secondary space hearing
5	Fuel used by built-in electric units for secondary space heating
5	Heating stove used for secondary space heating
5	Fuel used by heating stove for secondary space heating
5	Portable electric heaters used for secondary space heating
5	Portable kerosene heaters used for secondary space heating
5	Fireplace used for secondary space heating
5	Fuel used by fireplace for secondary space heating
5	Flue on gas fireplace
1	Frequency gas fireplace used
5	Cooking stove used for secondary space heating
5	Fuel used by cooking stove for secondary space heating
5	Other equipment used for secondary space heating
5	Fuel used by other secondary space heating equipment
6	Portion of space heating provided by main space heating equipment
5	Number of rooms heated
5	Thermostat(s) for heating equipment
5	Number of thermostats
5	Programmable main thermostat
1	Programmable thermostat lowers temperature at night
1	Programmable thermostat lowers temperature during the day
1	Temperature when someone is home during the day (winter)
1	Temperature when no on is home during the day (winter)
1	Temperature at night (winter)
1	Humidifier used
1	Number of months humidifier used in 2009
6	Number of tankless water heaters
6	Number of storage water heaters
6	Type of main water heater
6	Fuel used by main water heater
6	Main water heater is used by more than one housing unit
6	Main water heater size (if storage tank)
7	Main water heater age
6	Blanket around the main water heater (if storage tank)
7	Assistance for purchasing the water heater blanket
7	Year of the assistance for purchasing the water heater blanket
6	Type of secondary water heater
6	Fuel used by secondary water heater
6	Secondary water heater size (if storage tank)
6	Secondary water heater age
3	Air conditioning equipment used
3	No air conditioning equipment, or unused air conditioning equipment
5	Type of unused air conditioning equipment
5	Type of air conditioning equipment used
5	Ducts for space heating and air conditioning
5	Central air conditioner is a heat pump
5	Central air conditioner cools other homes, business, or farm

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
7	Routine service or maintenance performed on central air conditioner
7	Age of central air conditioner
7	Central air conditioner replaced by this household in the last 4 years
7	Assistance for maintaining or replacing central air conditioner
7	Year of assistance for central air conditioner
5	Number of rooms cooled
1	Frequency central air conditioner used in summer 2009
5	Thermostat for central air conditioner
5	Programmable thermostat for central air conditioner
1	Programmable thermostat adjusts temperature at night
1	Programmable thermostat adjusts temperature during the day
1	Temperature when someone is home during the day (summer)
1	Temperature when no on is home during the day (summer)
1	Temperature at night (summer)
5	Number of window/wall air conditioning units used
7	Age of most-used window/wall air conditioning unit
7	Energy Star most-used window/wall air conditioning unit
7	Most-used window/wall air conditioning unit replaced by this household in the last 4 years
7	Assistance for most-used window/wall air conditioning unit
7	Year of assistance for most-used window/wall air conditioning unit
1	Frequency most-used window/wall air conditioning unit used in summer 2009
5	Number of ceiling fans used
1	Frequency most-used ceiling fan used in summer 2009
5	Housing unit shaded from sun by large trees
1	Dehumidifier used
1	Number of months dehumidifier used in 2009
5	High ceilings
5	Cathedral ceilings
3	Swimming pool
3	Heated swimming pool
6	Fuel used for heating swimming pool
3	Hot tub used
6	Fuel used for heating hot tub
1	Number of lights turned on 12 or more hours during a typical summer day
2	Number of energy-efficient bulbs for lights turned on 12 or more hours during a typical summer day
1	Number of lights turned on 4 to 12 hours during a typical summer day
2	Number of energy-efficient bulbs for lights turned on 4 to 12 hours during a typical summer day
1	Number of lights turned on 1 to 4 hours during a typical summer day
2	Number of energy-efficient bulbs for lights turned on 1 to 4 hours during a typical summer day
1	Number of outdoor lights left on all night
2	Number of energy-efficient bulbs for outdoor lights left on all night
2	Number of outdoor lights left on all night that use natural gas
6	Energy-efficient bulbs installed by this household
7	Assistance for energy-efficient light bulbs
7	Year of assistance for energy-efficient light bulbs
5	Sliding glass doors in heated areas
6	Number of sliding glass doors in heated areas
6	Number of windows in heated areas
6	Type of glass in most windows
6	Windows replaced by this household
7	Assistance for window replacement

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
7	Year of assistance for window replacement
6	Level of insulation (respondent reported)
6	Insulation added by this household
6	Year of added insulation
7	Assistance for added insulation
7	Year of assistance for added insulation
6	Is home too drafty in the winter? (respondent reported)
6	Caulking or weather stripping by this household
6	Year of caulking or weather stripping
7	Assistance for caulking or weather stripping
7	Year of assistance for caulking or weather stripping
6	Home energy audit
6	Year of home energy audit
7	Assistance for home energy audit
7	Year of assistance for home energy audit
3	Electricity is used in home
3	Natural gas is used in home
3	Propane is used in home
3	Fuel oil is used in home
3	Kerosene is used in home
3	Wood is used in home
3	Solar is used in home
3	Other fuel is used in home
3	Electricity used for space heating
3	Electricity used for secondary space heating
3	Electricity used for air conditioning
3	Electricity used for water heating
3	Electricity used for cooking
3	Electricity used, other than for space heating, water heating, air conditioning, or cooking
3	Natural gas used for space heating
3	Natural gas used for secondary space heating
3	Natural gas used for water heating
3	Natural gas used for cooking
3	Natural gas used, other than for space heating, water heating, or cooking
3	Propane used for space heating
3	Propane used for secondary space heating
3	Propane used for water heating
3	Propane used for cooking
3	Propane used, other than for space heating, water heating, or cooking
3	Fuel oil used for space heating
3	Fuel oil used for secondary space heating
3	Fuel oil used for water heating
3	Fuel oil used, other than for space heating or water heating
3	Kerosene used for space heating
3	Kerosene used for secondary space heating
3	Kerosene used for water heating
3	Kerosene used, other than for space heating or water heating
3	Wood used for space heating
3	Wood used for secondary space heating
3	Wood used for water heating
3	Wood used, other than for space heating or water heating

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
3	Solar used for space heating
3	Solar used for secondary space heating
3	Solar used for water heating
3	Solar used, other than for space heating or water heating
3	Other fuel used for space heating
3	Other fuel used for secondary space heating
3	Other fuel used for water heating
3	Other fuel used for cooking
5	Renewable on-site system used
6	Renewable on-site system connected to the grid
3	Who pays for electricity used for space heating
3	Who pays for electricity used for water heating
3	Who pays for electricity used for cooking
3	Who pays for electricity used for air conditioning
3	Who pays for electricity used for lighting and other appliances
3	Follow up for 'other' payment of electricity
3	Who pays for natural gas for space heating
3	Who pays for natural gas for water heating
3	Who pays for natural gas for cooking
3	Who pays for natural gas for other uses
3	Follow up for 'other' payment of natural gas
3	Who pays for fuel oil
3	Follow up for 'other' payment of fuel oil
3	Who pays for propane
3	Follow up for 'other' payment of propane
3	Propane delivered
3	Kerosene delivered to home
3	Kerosene purchased 'cash and carry'
3	Wood logs used
3	Wood scraps used
3	Wood pellets used
3	Type of wood used other than logs, pellets, or scraps
7	Cords of wood used in 2009
7	Cords of wood used in 2009 (if more than 5)
1	Household fuel bills include fuel used for non-household purposes
1	Sex of householder
1	Employment status of householder
1	Householder lives with spouse or partner
1	Householder is Hispanic or Latino
1	Householder's Race
1	Highest education completed by householder
1	Number of household members
1	Age of householder
1	Age category of second household member
1	Age category of third household member
1	Age category of fourth household member
1	Age category of fifth household member
1	Age category of sixth household member
1	Age category of seventh household member
1	Age category of eighth household member
1	Age category of ninth household member

Continued on next page

Table B.1 – RECS feature cost categories, continued from previous page

Cost	Feature
1	Age category of tenth household member
1	Age category of eleventh household member
1	Age category of twelfth household member
1	Age category of thirteenth household member
1	Age category of fourteenth household member
1	Home-based business or service
1	Household member at home on typical week days
1	Household member(s) telecommutes or teleworks
1	Number of telecommuting days per month
1	Any activities that use an unusual amount of energy
1	Household members received employment income in 2009
1	Household members received retirement income in 2009
1	Household members received Supplemental Security income in 2009
1	Household members received welfare payments or cash assistance in 2009
1	Household members received investment income in 2009
1	Household members received other regular income in 2009
2	2009 gross household income
2	Household income at or below 100% of poverty line
2	Household income at or below 150% of poverty line
3	Housing unit in public housing authority
3	Lower rent due to Federal, State, or Local housing program
1	Household receives food stamps or WIC assistance
4	Total square footage (includes all attached garages, all basements, and finished/heated/cooled attics)
6	Total square footage (includes heated/cooled garages, all basements, and finished/heated/cooled attics)
6	Total heated square footage
6	Total unheated square footage
6	Total cooled square footage
6	Total uncooled square footage

B.2 World Values Survey

The results presented in this thesis come from the Wave 6 of the World Values Survey (WVS), available online[3].

We determined the cost, defined as item response times, for the WVS questions used in the secular and emancipative values scales (Tables B.2 and B.3) by conducting an experiment with 100 crowdworkers on Amazon Mechanical Turk.

Table B.2: Question costs (response times) for secular values scale on the World Values Survey.

Question	**Response time (s)**
Make parents proud	10.98
Greater respect for authority	16.41
Proud of nationality	6.54
Importance of religion	6.17
Religious service attendance	8.07
Religious person	6.30
Justifiable: avoiding fare	13.24
Justifiable: cheating on taxes	7.37
Justifiable: accepting a bribe	7.87
Confidence in armed forces	8.14
Confidence in police	7.42
Confidence in courts	5.43

Table B.3: Question costs (response times) for emancipative values scale on the World Values Survey.

Question	**Response time (s)**
Independence impt for children	6.53
Imagination impt for children	7.12
Obedience impt for children	5.79
Gender eq: job	9.52
Gender eq: education	8.35
Gender eq: political leader	6.61
Justifiable: homosexuality	8.71
Justifiable: abortion	6.55
Justifiable: divorce	6.69
Aims: say in jobs	37.92
Aims: say in gov't decisions	16.12
Aims: freedom of speech	10.41

[3] `www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`

B.3 SIPP Synthetic Beta

The results presented in this thesis come from Version 6.02 of the SIPP Synthetic Beta (SSB). The SSB is available for research upon application to the U.S. Census Bureau[4].

We determined the cost, defined as item nonresponse rates, for items in the SSB from publicly available SIPP data[5]. Items in the SSB that come from administrative records data (from the Social Security Administration) rather than the SIPP questionnaire have item nonresponse rates of 0. Table B.4 lists the cost for each SSB item.

Table B.4: Nonresponse rates for items in the SIPP Synthetic Beta. Some items come from administrative records from the Social Security Administration, which has information on respondents' earnings, benefits, *etc.* These items have zero cost, since respondents do not need to provide answers in the SIPP survey.

Question	**Nonresponse rate (%)**
Age	0
Respondent has valid Social Security Number	0
Home equity	0
Eligible for aged spouse benefit	0
Eligible for retirement benefit	0
Eligible for widowed spouse benefit	0
Non-housing financial wealth	0
Received aged spouse benefit	0
Received retirement benefit	0
Received widowed spouse benefit	0
Received Supplementary Security Income (SSI) benefits	0
Diagnostic code used to determine eligibility for SSI disability benefits	0
Type of SSI benefit received	0
Total earnings	0
Total personal income	0
Gender	0.002
Received workers compensation	0.026
Own a home	0.030
Received public assistance payments	0.031
Disability prevents work	0.038
Received veterans compensation or pension benefits	0.072
Received SNAP / food stamps	0.136
Enrolled in defined benefit pension plan	0.279
Weeks with pay	1.035
Occupation category	1.339
Education category	1.364
Field of bachelors degree	1.598
Race	1.813
Has kids	2.511
Health insurance coverage	3.170
Hispanic	3.202
Life insurance ownership	3.640
Enrolled in defined contribution pension plan	4.104
Disability limits work	5.515
On layoff (without pay)	5.956
Left a job because of layoff	5.956
Total usual weekly hours worked at all jobs	6.379

Continued on next page

[4] https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html
[5] http://thedataweb.rm.census.gov/ftp/sipp_ftp.html

Table B.4 – SSB item costs, continued from previous page

Question	Nonresponse rate (%)
Health insurance coverage from employer	6.708
Foreign born	6.929
Currently enrolled in college	7.549
Currently enrolled in high school	8.614
Industry category	17.440

B.4 American Community Survey

The results presented in this thesis come from data (responses and paradata) collected June through September 2016, on the online mode of the American Community Survey (ACS).

We determined the cost, defined as time (in seconds) that respondents spent on a page in the online ACS questionnaire. Table B.5 lists the cost for each page in the ACS.

Table B.5: Median item response times for pages in the online ACS.

Page type	Page	Median time on page (s)
login	login	8
setup	address	6
setup	liveu	5
setup	live	6
setup	business	6
setup	pin	35
setup	resp_name	26
setup	roster_a	30
setup	roster_b	10
setup	add_1	18
setup	roster_c	9
setup	add_2	18
setup	away_now	10
setup	remove_one	7
setup	another_home	7
setup	another_home_who	7
setup	more_than_2	6
setup	roster_check	5
setup	ref_per	13
roster_demos	relationship	6
roster_demos	sex	3
roster_demos	dateofbirth	18
roster_demos	hispanic	4
roster_demos	race	6
household	typeofunit	16
household	yearbuilt	13
household	whenmovedin	19
household	acres	5
household	agrsales	8
household	rooms	42
household	facilities	15
household	compuse	15

Continued on next page

Table B.5 – ACS page times, continued from previous page

Page type	Page	Median time on page (s)
household	netaccess	12
household	netsub	20
household	vehicles	9
household	heatingfuel	12
household	elecpay	8
household	elecamt	21
household	elecinc	14
household	gasuse	5
household	gaspay	7
household	gasamt	10
household	gasinc	12
household	waterpay	5
household	wateramt	26
household	waterinc	8
household	ofueluse	6
household	ofuelpay	8
household	ofuelamt	16
household	ofuelinc	9
household	foodstamps	7
household	condo	4
household	condofee	7
household	condofeeamt	9
household	tenure	11
household	monthrent	8
household	meals	4
household	propvalue	18
household	taxes	17
household	propinsurance	17
household	mortgage	12
household	mortgageamt	14
household	mortgagetax	8
household	mortgageinsurance	6
household	2ndmortgage	7
household	2ndmortgageamt	11
household	mobilehometax	31
household	hunitstatus	21
person	pselect	7
person	placeofbirth	12
person	citizenship	13
person	yearofentry	12
person	attendschool	7
person	whatgrade	8
person	highestlevel	13
person	fieldofdegree	17
person	ancestry	15
person	language	4
person	whatlanguage	6
person	englishprof	5
person	residencelastyear	5
person	addresslastyear	20
person	insurance	18

Continued on next page

Table B.5 – ACS page times, continued from previous page

Page type	Page	Median time on page (s)
person	deaf	4
person	blind	4
person	difficultyconcent	5
person	difficultywalk	4
person	difficultydress	3
person	difficultyerrand	5
person	marriedstatus	5
person	pmarried	4
person	widow	4
person	divorce	3
person	numberofmarriages	4
person	yearofmarriage	9
person	birth	4
person	grandchildrenhome	5
person	grandparentsresp	7
person	lengthofresp	9
person	veteranstat	5
person	periodofservice	13
person	vadisability	6
person	disabilityrate	4
person	worklastweek	6
person	worklocal	55
person	transporttowork	6
person	numberofriders	7
person	timeleftforwork	14
person	mintowork	8
person	fiftymoreweeks	9
person	weeksworked	13
person	hoursworked	10
person	anywork	6
person	layoff	4
person	tempabsent	5
person	recalltowork	4
person	activelookforwork	5
person	couldwork	9
person	lastworked	9
person	employeetype	18
person	employer	13
person	militaryemployer	4
person	typeofbusiness	12
person	businessclass	7
person	typeofwork	11
person	duties	16
person	wages	10
person	wagesamt	20
person	selfemp	6
person	selfempamt	15
person	interest	7
person	interestamt	21
person	socialsecurity	5
person	socialsecurityamt	37

Continued on next page

Table B.5 – ACS page times, continued from previous page

Page type	**Page**	**Median time on page (s)**
person	ssi	4
person	ssiamt	16
person	publicasst	4
person	publicasstamt	15
person	retirement	4
person	retirementamt	26
person	otherincome	6
person	otherincomeamt	20
person	totalincome	8
person	vrfyincome	7
person	estincome	14
person	finishedperson	4

Appendix C

Demographic characteristics and breakoff statistics on the online ACS

Breakoff patterns by respondent demographic group in the online ACS agree with most trends found in previous work. One demographic characteristic of breakoffs that was not reproduced in this ACS dataset is the past finding that men are more likely to break off than women (Peytchev, 2011). Table C.1 shows the numbers of breakoff and complete responses from women and men—slightly more women than men were respondents for the survey (52.20% of respondents were women), and women broke off at a higher rate than men (12.37% breakoff for women, 9.78% breakoff for men). Remaining demographics and breakoff were observed in accordance with prior work. Table C.2 shows that older respondents completed the survey at higher rates than younger respondents (McCutcheon et al., 2013; Peytchev, 2009, 2011). Tables C.4 and C.3 show that nonwhite and Hispanic respondents, respectively, were more likely to break off than white and non-Hispanic respondents (Peytchev, 2011). Table C.5 shows that respondents with higher education levels completed the survey at greater rates than respondents with lower education levels (Peytchev, 2009). Finally, Table C.6 shows that low-income respondents were more likely to break off than higher-income respondents (Peytchev, 2009; Goyder et al., 2002; Van Goor & Rispens, 2004).

Table C.1: Counts of respondent sex by completion type. Percentages of each sex that are breakoffs or completes are in parentheses. More women than men were respondents, and women broke off at a higher percentage than men.

Sex	**Breakoff**	**Complete**
Missing	503 (78.35)	139 (21.65)
Male	15120 (9.78)	139600 (90.22)
Female	21000 (12.37)	148700 (87.63)
Total	36630 (11.27)	288400 (88.73)

Table C.2: Counts of respondent age by completion type. Percentages of each age category that are breakoffs or completes are in parentheses. Older respondents completed the survey at higher rates than younger respondents.

Age	**Breakoff**	**Complete**
Missing	2504 (64.84)	1358 (35.16)
<18	135 (31.54)	293 (68.46)
18-34	7926 (12.52)	55400 (87.48)
35-49	11060 (12.43)	77920 (87.57)
50-64	10110 (9.84)	92680 (90.16)
65+	4893 (7.45)	60770 (92.55)
Total	36630 (11.27)	288400 (88.73)

Table C.3: Counts of respondent Hispanic origin by completion type. Percentages of each category that are breakoffs or completes are in parentheses. Non-Hispanic respondents completed the survey at higher rates than Hispanic respondents.

Hispanic origin	**Breakoff**	**Complete**
Missing	3304 (71.58)	1312 (28.42)
Hispanic	4276 (17.43)	20260 (82.57)
Not Hispanic	29050 (9.82)	266900 (90.18)
Total	36630 (11.27)	288400 (88.73)

Table C.4: Counts of respondent race by completion type. Percentages of each race that are breakoffs or completes are in parentheses. White respondents completed the survey at higher rates than nonwhite respondents.

Race	**Breakoff**	**Complete**
Missing	3693 (67.32)	1793 (32.68)
Not white	6738 (14.63)	39310 (85.37)
White	26200 (9.58)	247300 (90.42)
Total	36630 (11.27)	288400 (88.73)

Table C.5: Counts of respondent education by completion type. Percentages of each education level that are breakoffs or completes are in parentheses. Respondents with higher education levels completed the survey at greater rates than respondents with lower education levels.

Education	**Breakoff**	**Complete**
Missing	10480 (92.01)	910 (7.99)
≤HS	6461 (10.30)	56260 (89.70)
Some college	9114 (9.22)	89700 (90.78)
Bachelors	6455 (7.36)	81210 (92.64)
Grad school	4125 (6.40)	60350 (93.60)
Total	36630 (11.27)	288400 (88.73)

Table C.6: Counts of respondent income category by completion type. Percentages of each income category that are breakoffs or completes are in parentheses. Respondents with higher incomes completed the survey at greater rates than lower-income respondents.

Income ($)	**Breakoff**	**Complete**
Missing	26000 (63.78)	14760 (36.22)
Loss	37 (5.62)	622 (94.39)
0	1079 (6.50)	15510 (93.50)
1-19,999	2805 (5.10)	52150 (94.90)
20,000-49,999	3128 (3.54)	85150 (96.46)
50,000-69,999	1386 (3.12)	42960 (96.87)
70,000-99,999	1079 (2.98)	35180 (97.02)
100,000+	1122 (2.60)	42100 (97.40)
Total	36630 (11.27)	288400 (88.73)