# The Multidimensional Nature of Intrahousehold Power

Rossella Calvi[*] Sugat Chaturvedi[†] Jacob Penglase[‡]

November 2025

**Abstract**

Measuring intrahousehold power is challenging due to limited data and the complexity of household dynamics. Power also manifests differently across decision domains. Using rich data from married couples in Bangladesh and three machine learning methods, we develop an empirical framework to capture its multidimensional nature. We examine how well standard power-related survey questions predict relative spousal wellbeing in time use, consumption, and health. Indicators that are strong in one domain often perform poorly in others, underscoring the domain-specific nature of power. To interpret these patterns, we extend the collective household model to allow for domain-specific bargaining weights under sequential decision-making.

**JEL Codes**: C83, D13, I31 J16, O12.

**Keywords**: intrahousehold power, bargaining, collective model, machine learning, feature selection.

[*]Rice University, Department of Economics, 6100 Main Street, Houston TX, 77005 (e-mail: rossella.calvi@rice.edu).

[†]School of Public Policy, Indian Institute of Technology Delhi, IIT Campus, Hauz Khas, Delhi 110016, India (e-mail: sugatc@iitd.ac.in).

[‡]San Diego State University, Department of Economics, 5500 Campanile Drive, San Diego CA, 92182 (e-mail: jpenglase@sdsu.edu).

We thank Samson Alva and seminar participants at the UCSD/UCLA/CIFAR Conference on Economic Development, Political Economy, and Culture. All errors are our own.

# 1 Introduction

Measuring decision-making power and welfare within households is challenging. Household surveys often collect data at the household rather than individual level, making it difficult to observe intrahousehold dynamics or the decision-making process between spouses. This task is further complicated by intrahousehold power spanning several spheres, from freedom of mobility and agency in time allocation to control over family resources.

Social scientists have proposed multiple ways to measure women's relative power and agency within families. These include widely used survey-based indices, such as the Women's Empowerment in Agriculture Index (WEAI; Alkire *et al.* (2013)) and the Survey-based Women's Empowerment Index (SWPER; Ewerling *et al.* (2017)). More recent approaches draw on machine learning and qualitative tools, such as the Machine Learning and Semi-structured Interviews (MASI) method of Jayachandran *et al.* (2023). Lab-in-the-field experiments have also been employed to improve measurement by eliciting willingness to pay for control over resources as a proxy for women's decision-making power (Almås *et al.*, 2018). Beyond empirical methods, theory-based approaches such as the collective household model (Chiappori, 1988, 1992; Browning *et al.*, 2013) provide a systematic framework for conceptualizing and analyzing women's power within the family. While each of these approaches has advanced the field, most either emphasize specific domains of power or aggregate diverse dimensions into a single index, potentially obscuring the multidimensional nature of intrahousehold power.[1]

Our study examines the multidimensional nature of intrahousehold power by evaluating how well questions from commonly used survey modules predict different aspects of women's wellbeing relative to their husbands.[2] Specifically, we investigate whether indicators of power in one domain (e.g., consumption) are also good predictors in other domains (e.g., health and time use). Our primary objective is not to propose new survey instruments but to assess whether existing survey questions capture a unified dimension of power or instead reflect distinct, domain-specific aspects with heterogeneous policy implications for improving individual wellbeing.

We draw on a comprehensive household survey from Bangladesh that combines rich questionnaires on women's and men's intrahousehold power with detailed measures of individual consumption, health, and time use. The survey includes questions on a wide range of topics, from participation in household economic decisions to experiences of domestic abuse. Importantly, many of these questions were asked of both husbands and wives, allowing us to assess not only which questions are most informative but also how responses and their predictive performance differ by respondent.

We begin by applying machine learning methods to examine the multidimensional nature of intrahousehold power. We posit two possibilities. The first is that power is unidimensional: strength in one domain, such as decision-making over time use, would reliably reflect strength in all others, such as control over household resources or agency over one's own health. In this case, the same survey questions on intrahousehold power should consistently predict relative wellbeing across diverse dimensions. The second possibility is that power is multidimensional: bargaining power in

[1] One exception is the Women's Empowerment in Agriculture Index (WEAI), which is explicitly multidimensional, encompassing five domains: production, resources, income, leadership, and time (Alkire *et al.*, 2013; Quisumbing, 2025).

[2] While women's power extends beyond the household—political participation, representation, and broader societal influence—our analysis focuses on intrahousehold power and wellbeing.

one sphere does not necessarily translate into bargaining power in another, so different aspects of wellbeing would be best captured by distinct sets of survey questions rather than a single overarching measure.[3]

Our results align with the multidimensional view. Questions that strongly predict one outcome (such as relative time use) are often weak predictors of others (such as relative consumption or health). This pattern suggests that intrahousehold power operates through multiple, domain-specific mechanisms, challenging the notion of a singular measure and underscoring the need to tailor measurement tools to the particular dimension under study.

To identify the survey questions most predictive of intrahousehold relative outcomes in various dimensions, we employ a range of machine learning techniques. Our preferred approach utilizes the random forest algorithm, which ranks questions based on how informative their responses are for predicting multiple correlated target variables. For instance, when analyzing time use, we consider several related outcomes: women's market work, household work, and stated satisfaction with time use relative to their husbands. Similarly, for food consumption, we include both women's caloric and micronutrient intakes and the monetary value of food consumed relative to their spouses. For health status, we consider measures of both short-term and long-term illnesses. To ensure robustness, we also apply alternative machine learning methods—such as multi-task Lasso and group Lasso stability selection—to assess the sensitivity of our findings to the choice of prediction algorithm.

Our results show that the survey questions whose responses are the most predictive of intrahousehold relative outcomes vary substantially across domains. Based on the random forest algorithm, we find little to no correlation in feature importance rankings across time use, consumption, and health. For example, the correlation between feature importance in predicting relative time allocation and health is just 0.0935; the correlation between the predictive performance of responses to power-related survey questions for consumption and health is similarly low. Results from group Lasso and multi-task Lasso stability selection further reinforce this finding: among the top 10 survey questions for each outcome, there is no overlap across the three dimensions. Even when expanding to the top 50 questions, overlap remains minimal, with only 14 percent of questions identified as top predictors in more than one domain.

We conclude the paper by linking our empirical findings to a simple conceptual framework that formalizes intrahousehold power as a multidimensional construct. Building on the formulation of the collective household model of Chiappori (1988, 1992), Browning *et al.* (2013) and Dunbar *et al.* (2013), our model allows for domain-specific bargaining weights and recognizes that decisions in one domain may influence both choice sets and intrahousehold bargaining power concerning other issues. Each issue is resolved in sequence and is not revisited in the short term, reflecting frictions such as coordination costs, limited attention, or limited foresight. This framework implies that household decisions in one domain may be efficient given choices in other domains, but household decisions may not necessarily be efficient overall (Lewbel & Pendakur, 2022). Importantly, the model highlights that intrahousehold bargaining power may be multidimensional and issue-specific, challenging the common assumption that a single index can summarize the intrahousehold decision process.

[3] Simple correlations of relative outcomes or power-related questions across domains offer only a blunt summary of how domains co-move. For instance, they cannot show which survey questions actually carry cross-domain signal, how much of that signal is shared or domain-specific, or whether relationships run more strongly in some directions than others. Our approach provides a structured way to uncover these patterns.

Our main contribution is to incorporate machine learning techniques to study the measurement of intrahousehold power and its relationship to relative wellbeing.[4] The focus on the multidimensionality of women's power builds on Kabeer (1999), who emphasizes the need for comprehensive measures capturing women's access to resources, power, and achievements, consistent with Sen's capabilities framework (Sen, 1990; Sen *et al.*, 1999). Laszlo *et al.* (2020) similarly argue that traditional approaches—centered on income and asset ownership—miss the complex dynamics of decision-making within households. Donald *et al.* (2020) make related points and propose a multidisciplinary framework to address these gaps. Relatedly, Lundberg & Pollak (1993) formalize a model where responsibilities and power are divided into separate spheres, shaped by social norms.

We also contribute to recent work on developing survey tools to measure women's power. Jayachandran *et al.* (2023) combine machine learning with qualitative measures to design a concise survey module. Our goals differ: rather than designing new instruments, we assess whether existing questions predict different observable aspects of women's wellbeing—specifically time use, consumption, and health—rather than qualitative measures from semi-structured interviews.

The policy relevance of *unbundling* women's power is highlighted by Anderson (2022), who stresses that its dimensions interact and co-evolve in ways not yet fully understood. Policies often assume that greater power leads to broader wellbeing gains. Evaluating how well survey questions predict outcomes across domains is therefore essential. Our paper takes a step in this direction through both measurement and conceptual contributions.

# 2 Data

We use data from the 2018–2019 Bangladesh Integrated Household Survey (BIHS), conducted by IFPRI. The survey includes rich measures of intrahousehold power alongside detailed information on health, individual consumption, and time use. We focus on 1,620 rural households with married couples (with or without children), excluding those with multiple adult men or women to avoid more complex family structures.

## 2.1 Outcomes

Our objective is to predict women's wellbeing relative to their spouses, encompassing multiple dimensions that are often challenging to measure. The design of BIHS allows us to address these challenges, as it collects detailed information on time use, health, and individual consumption for both women and men. Table A2 in the Appendix lists the outcome measures used in our analysis, which focus on these three key dimensions of welfare.[5]

Time use is measured in hours spent on income-generating and domestic work. Respondents completed a 24-hour diary, recording activities in 15-minute intervals across 23 categories such as "sleeping," "personal care," and "farming." For analysis, these are aggregated into leisure, domestic,

[4] Jayachandra & Voena (2025) provide an insightful review of both the measurement and theory of women's intrahousehold power in low-income settings.

[5] In Section 4, we assess how well power-related survey questions predict relative intrahousehold wellbeing by reporting $R^2$ from regressions of relative outcomes on the top predictors identified through machine learning algorithms. As a benchmark, we also regress each outcome on standard demographic characteristics commonly available in household rosters (see the rightmost column of Table A2).

and income-generating work. Alongside these objective measures, we construct an index of subjective satisfaction based on survey questions asking respondents to rate, from 1 ("not satisfied") to 10 ("very satisfied"), aspects such as the division of household work and leisure time. We combine these ratings using principal component analysis to create a composite satisfaction index.

In our sample, households allocate over two-thirds of their budget to food, on average. We therefore use relative food consumption as a meaningful proxy for relative intrahousehold consumption. Food quantity is assessed through caloric intake and the monetary value of total food consumption, while food quality is evaluated on the basis of nutrient intake, including protein, iron, vitamins, and zinc. These measures are constructed from a detailed survey module that recorded individual food consumption over the course of a single day. To generate these data, enumerators interviewed the household member primarily responsible for food preparation, collecting information not only on the meals cooked and ingredients used, but also on which household members consumed each meal.[6]

To construct health outcomes, we use principal component analysis to summarize survey questions on long-term disabilities and short-term morbidities. The first set captures functional limitations—such as difficulties standing, walking long distances, carrying loads, or with hearing, speech, and eyesight—reflecting chronic conditions that affect participation in daily activities. The second set focuses on recent illnesses (past four weeks), including fever, diarrhea, and cough, as well as diagnosed conditions like malaria or pneumonia, alongside workdays lost. We also compute BMI as a complementary health indicator.[7] Finally, we construct an index of domestic abuse based on reports of threats of divorce, and verbal or physical violence.

Since our analysis focuses on intrahousehold relative wellbeing between spouses, we transform the outcome variables described above—with the exception of the domestic violence index, which is not measured for men—by dividing women's values by the sum of women's and men's values. The distributions of these relative outcomes across different domains are shown in Figure A2 in the Appendix.

## 2.2 Predictors

In Table A1 in the Appendix, we summarize the survey modules used to measure relative intrahousehold power and the number of questions drawn from each. These questions serve as possible predictors of relative spousal outcomes discussed above. The "women's status" component captures the degree to which women are allowed to work or face opposition from husbands or in-laws. It also records whether women retain, share, or surrender control over their earnings (if any), how decisions on everyday expenses or loans are made, and whether women themselves can make such financial decisions. It also tracks freedom of movement to markets, clinics, or social events, as well as autonomy in reproductive choices. The WEAI module complements this information by shifting the focus to economic and civic agency more broadly. It probes women's decision roles in farming, livestock, business, and wage employment, and whether they can influence how income and assets are used. Ownership and control over land, equipment, housing, and other durable goods are

[6] Several precautions were taken to enhance reliability and reduce measurement error. Details in D'Souza & Tandon (2019) and Brown *et al.* (2021).

[7] BMI can also be viewed as a nutrition outcome. As a robustness check, we classify it under food quantity and find similar results.

documented alongside access to loans and household spending.[8]

Note that the sets of predictors by design differ by gender. While the majority of questions are asked to both men and women (that is, to both spouses within the same couple), certain modules, such as the one focused on freedom of mobility, are administered only to women. Further differences in the number of available predictors arise because questions with no variation in responses are excluded from the analysis. In specifications that pool men's and women's predictors, a predictor is an individual response, which is essentially defined as the interaction between a survey question and the sex of the respondent.[9]

All questions in these modules are close-ended and include both binary and multiple-choice formats. Some questions require a simple "yes" or "no" response, such as "Do you alone have any money that you can decide how to spend?" Others present multiple options, such as "Who made the decision to use birth control?" with possible answers including "yourself," "your husband," "both yourself and your husband," or "someone else." For questions with categorical responses that lack an inherent ranking, we re-code the answers into multiple indicator variables. Consequently, the survey questions translate into a total of 1,078 predictor variables across men and women.[10,11]

# 3 Methodology

This section outlines our approach to selecting survey questions that best predict spousal relative outcomes. We frame the task as a multi-task prediction problem.[12] While main results use the pooled set of women's and men's questions, we also estimate models separately by respondents' gender.

## 3.1 Random Forest

Our preferred approach employs a random forest regressor, an ensemble learning method that constructs multiple decision trees on bootstrapped samples of the training data. Decision trees partition the data by recursively applying decision rules that minimize mean squared error (i.e., within-node variance). Because individual trees tend to overfit, random forests introduce two sources of randomness—bootstrapping the data and selecting a random subset of features at each split—which improves generalization.

[8] In addition, the WEAI module includes questions that may seem only peripherally related to intrahousehold power. In particular, it asks respondents to reflect on the motivations behind their actions in multiple domains—such as work, household expenditures, health, religion expression, and family planning—and to rate the extent to which their behavior is driven by external incentives (punishment or rewards), social expectations (reputation, avoiding blame), or personal values and interests. It also records satisfaction with decision-making in these areas.

[9] Prior research by Ghuman *et al.* (2006) and Ambler *et al.* (2021) document significant discrepancies between spouses in their responses to questions about empowerment. The findings of Ambler *et al.* (2021) are particularly relevant, as their study also utilizes the BIHS and reveals substantial disagreement regarding decision-making authority and asset ownership within households.

[10] Specifically, the 172 survey questions yield 591 predictor variables for women, while the 143 questions yield 487 predictor variables for men. In certain specifications, we restrict our algorithm to select questions (i.e., groups of predictors) rather than individual predictors. This is discussed further in Section 3.

[11] Not all survey questions are asked of every respondent due to skip sequencing or conditional branching (see Manski & Molinari, 2008). For example, only landowners are asked who decides how land income is spent. To retain such questions without dropping non-landowners, we merge initial and follow-up questions into composite variables—for instance: "Do you own land, and if so, who decides how income is spent?" with response options capturing own, spouse, joint, or not applicable. All predictors are thus recoded for consistency and coverage.

[12] A multi-task prediction problem in machine learning is a setting where a model is trained to predict multiple related outcomes at the same time, rather than building a separate model for each outcome. Further technical details are in Appendix A.

To assess the predictive performance of power-related survey questions, we use permutation feature importance (Breiman, 2001), which measures the decrease in predictive accuracy when a given feature is randomly shuffled—features whose shuffling leads to a larger drop in accuracy are those with greater predictive power. This approach mitigates biases in impurity-based importance and yields robust rankings. We repeat the procedure 100 times on random 50 percent subsamples of the data and use five-fold cross-validation to tune hyperparameters. The most relevant features are ranked by their average decrease in predictive accuracy across iterations.

### 3.2 Alternative Approaches

We test the robustness of our findings to the use of different machine learning algorithms. In an alternative approach, we employ multi-task Lasso (Argyriou *et al.*, 2006), an extension of the standard Lasso model that selects predictors for multiple related outcomes simultaneously. By enforcing a shared sparsity constraint across tasks, this method ensures that the same set of features is selected for different related outcomes. To enhance the robustness of feature selection, we integrate multi-task Lasso with stability selection (Meinshausen & Bühlmann, 2010), following an iterative subsampling approach. Specifically, we repeatedly apply multi-task Lasso to random 50 percent sub-samples of the data and use five-fold cross-validation to determine the optimal regularization parameter. This process is repeated over 1,000 iterations, and predictors are ranked based on their selection frequency.[13]

As discussed in Section 2, survey responses often in the form of categorical variables are converted into multiple binary indicators. To address this specific feature of the data, we also use group Lasso (Yuan & Lin, 2006) to ensure that entire questions, rather than individual binary indicators, are selected jointly. This method imposes a structured sparsity constraint at the group level by penalizing all coefficients belonging to the same question together, thereby enforcing group-wise selection. As with multi-task Lasso, we integrate group Lasso with stability selection, applying the method iteratively to sub-samples of the data. Predictors are then ranked based on their selection frequency.

## 4 Results

### 4.1 The Multidimensionality of Women's Power

We begin by ranking the power-related survey questions asked to women and men according to their feature importance in predicting relative intrahousehold outcomes across three dimensions: time allocation, consumption, and health status. These rankings are generated using the random forest algorithm described in the previous section. To assess whether the same set of questions consistently predicts women's relative wellbeing across domains, we compare importance scores across the three outcomes. Because the scores are not directly comparable in magnitude, we focus on the

[13] We conduct sensitivity checks using standard Lasso stability selection, which does not impose the restriction of selecting the same features across different outcomes within a dimension. Our results are confirmed and available upon request.

rank correlations between them.[14]

Figure 1 illustrates the correlation—or more accurately, the lack of correlation—in rankings across the three dimensions. Stated differently, the figure reveals that distinct sets of survey questions are driving variation in different aspects of women's outcomes relative to their husbands. If a single set of power-related questions were consistently predictive across all domains, we would expect the data points to align closely along the 45-degree line in each panel. Significant deviations from this pattern support our hypothesis that women's intrahousehold power is indeed not unidimensional.

In line with this insight, we find that pairwise correlation coefficients of importance scores across dimensions are very small, and in some cases even negative. The rank correlation between the importance of survey questions in predicting women's relative time allocation and health status is just 0.0935. Importantly, the correlation between the importance scores of predictors of relative consumption and health is negative at -0.0444, although not significant at conventional levels. Interestingly, we detect a negative and statistically significant relationship between the importance of survey questions in predicting women's relative time allocation and consumption, with a correlation coefficient equal to -0.1399.[15]

When comparing the highest-ranked survey questions across domains, we find strikingly little overlap. None of the top 10 predictors of intrahousehold relative outcomes in time allocation, consumption, and health are shared across domains. In fact, no single power-related question appears among the top 50 predictors in all three domains. Overlap is limited even pairwise—between time use and health, time use and consumption, or consumption and health. Expanding the set to the top 100 predictors yields only three questions common to all domains.[16]

A broadly consistent picture emerges when we turn to the results from the group Lasso and multi-task Lasso stability selection algorithms. Among the top 10 predictors of women's relative outcomes in time allocation, health, and consumption—defined by their frequency of selection across 1,000 iterations of stability selection—there is no overlap across domains under either method. As we expand the predictor set, some overlap begins to emerge, though it remains limited. For instance, among the top 50 predictors, only about 14 percent are identified as strong predictors in more than one domain by both the multi-task Lasso and the group Lasso. When we consider the top 100 predictors, the extent of overlap (whether complete or pairwise across domains) increases mechanically, as expected, but it is still modest.

Finally, we replicate the analysis restricting the set of predictors to survey questions asked exclusively to women. The results closely mirror those in the main text. Predictor rankings are again essentially uncorrelated across domains, with rank correlations hovering near zero—for instance, 0.0171 for time use versus health and 0.0498 for consumption versus health. These low correlations confirm that the questions most predictive in one domain carry little information about outcomes in

[14]It is important to note that these rank correlations reflect the similarity in the predictive power of survey questions across domains, not the correlation (or lack thereof) in the underlying relative outcomes themselves. For example, even if women's relative outcomes in health and time use were negatively correlated, we might still observe a positive correlation in the predictive power of certain survey questions if those questions are systematically informative for both domains. Moreover, our exercise does not assume women's relative preferences for one outcome—such as health or consumption—mirror those for another, such as time use. Rather, our prediction framework is agnostic about preferences and instead seeks to identify distinct domains of intrahousehold power.

[15]These results echo findings in Bayissa *et al.* (2018), who use principal components analysis on Ethiopian data to separately quantify women's empowerment in familial, legal, psychological, political, and socio-cultural dimensions, finding little correlation across the five.

[16]For completeness, we provide the list of ten most predictive survey questions across each of our three key welfare domains for both our women's and men's surveys. The results using random forest are presented in Tables A3, A4, and A5 in the Appendix.

another. Overlap across domains is likewise minimal: no common predictors among the top 10, just one among the top 50, and six among the top 100.

As discussed earlier, if intrahousehold power was unidimensional, we would expect a high degree of overlap—if not complete overlap—across domains in the set of power-related questions that predict women's relative wellbeing in different domains. Instead, our findings suggest that different dimensions of relative intrahousehold wellbeing are predicted by very distinct sets of questions, reinforcing the view that intrahousehold power may not be captured by a single measure.

### 4.2 Cross-Domain Predictive Patterns

We next classify the top 50 predictors from the random forest analysis into broader thematic groups, which we refer to as *power domains*. This aggregation allows us to assess whether power in a given domain predicts outcomes primarily within the same domain or also across others. To validate this classification, in Appendix C we compare our manual groupings with those obtained through text-embedding clustering and find strong alignment.

Figure 3 maps the 50 most predictive survey questions (left) to relative outcomes (right). Flows indicate the share of top predictors originating in each power domain. Less common domains—such as fertility, agriculture, savings, and mobility—are grouped under "Other." Two patterns emerge. First, while individual predictors differ across outcomes, questions from one domain often predict outcomes in others. For instance, 16 percent of top predictors for health come from the consumption domain, and 22 percent from time use. Note that this does not imply overlap in the specific predictors across dimensions; rather, some questions that conceptually belong to one domain are empirically predictive of outcomes in others. Second, own-domain predictors remain critical: consumption questions most strongly predict consumption, health questions health, and so on.

Together, these patterns provide empirical support for a multidimensional view of intrahousehold power. This has important implications for both measurement and policy design. From a measurement perspective, focusing on a single domain is unlikely to capture the full structure of power within households. From a policy perspective, interventions that strengthen agency in one domain may yield spillover effects in other dimensions of wellbeing. Additional analyses—presented in Appendix B—show that (i) women's self-reported answers are more predictive than men's but the two are complementary, and (ii) predictive power is concentrated in a relatively small number of survey items, with time use outcomes being most predictable.

## 5 Conceptual Framework

To rationalize our empirical findings, this section develops a simple theoretical framework that extends the collective household model to incorporate sequential decision-making across multiple domains. The framework allows for domain-specific bargaining weights, explicitly capturing the multidimensional nature of intrahousehold power we have discussed in the previous sections.[17]

[17] We wish to note that the recent formulations of Browning *et al.* (2013) focus primarily on household decision-making over consumption allocations. These models typically assume that decisions in other domains—such as labor supply, health investments, or time use—are separable from the consumption decision. As a result, the bargaining process and the analysis of efficiency are confined to the consumption space. In

## 5.1 A Multi-issue Collective Household Model

We model the household as a collection of individuals, each with their own preferences, who jointly solve a sequential decision problem across $S$ *issues* or *domains*. In each domain $s = 1, \ldots, S$, the household makes a joint decision on a continuous variable $f_s \in F_s$ (e.g., time allocation, consumption of particular goods, or health investment).

A key feature of this framework is that once a decision is made in a particular domain, it is not revisited or easily renegotiated. This irreversibility may stem from a range of frictions, including coordination costs, cognitive limitations, or institutional constraints that make collective decision-making across multiple domains burdensome. For example, negotiating every issue simultaneously may simply be too complex or overwhelming for household members, particularly when decisions span diverse aspects of life—such as consumption, labor supply, or household investments. In addition, imperfect foresight and limited attention may prevent individuals from anticipating how current decisions interact with or constrain choices in other domains.[18] As a result, decisions are often made with partial information about their broader implications, and once implemented, they are difficult to reverse in the near term.

Let $J$ denote the set of decision-makers (e.g., $J = 2$ when the husband and wife are the only decision-makers). In each decision domain $s$, the household solves the following optimization problem:

$$\max_{\{f_j^s\}_{j\in J}} \sum_{j\in J} \mu_j^s(p^s, y^s, \phi^{s-1}) \Big[ v_j(g_j^s) + u_j(f_j^s) \Big] \tag{1}$$

subject to household-level budget and time constraints, as well as technological constraints that capture household production and consumption technologies. Here, $f^s$ denotes the decision made in domain $s$, and $\phi^{s-1} = \{f^1, \ldots, f^{s-1}\}$ captures the sequence of decisions made in other domains. The vector $g^s$ summarizes realized outcomes in other domains that are relevant for current utility. While closely related to $\phi^{s-1}$, it need not be identical.[19]

The function $u_j(f_j^s)$ represents member $j$'s direct utility from the decision in domain $s$, while $v_j(g_j^s)$ captures their utility from allocations in other domains. Both $u_j(f_j^s)$ and $v_j(g_j^s)$ are assumed to be monotonically increasing, continuously twice differentiable and strictly quasi-concave utility functions. The Pareto weights $\mu_j^s(\cdot)$ reflect member $j$'s relative influence in domain $s$. Both the household technology constraints and the Pareto weights associated with domain $s$ may depend on the prior decisions in other domains. From a bargaining perspective, the Pareto weight can be seen as a measure of individuals' influence in the decision process. The larger the value of the Pareto weight for household member $j$, the greater the weight her preferences receive in the resulting household program, and the greater will be the resulting allocation for member $j$. Stated differently, the higher the Pareto weight associated with individual $j$ when deciding on domain $s$, the more favorable the

contrast, our framework explicitly models household decisions across multiple domains, recognizing that choices in one domain can shape both the distribution of bargaining power and the feasible set of outcomes in subsequent domains.

[18] In this regard, our simple framework departs from Basu (2006), who proposes a model where bargaining power depends endogenously on household decisions, such as labor supply. While the framework adopts cooperative bargaining, the dependence of threat points on outcomes introduces strategic incentives that may lead to Pareto-inefficient equilibria. For instance, spouses may over-supply labor to boost their bargaining weight, even when both would benefit from working less. For simplicity, we abstract from strategic considerations in our simple framework.

[19] Note that $g_s$ and $\phi^s$ may or may not perfectly coincide. Although $\phi^{s-1}$ records the decisions made in earlier domains, the actual outcomes relevant when the couple decides on domain $s$—captured by $g_s$—may reflect a combination of those earlier decisions and their downstream consequences.

outcome for individual $j$ in that domain.[20]

Note that when there is only one decision to be made (that is, when $S = 1$), or when all decisions across multiple domains are negotiated jointly rather than sequentially, the model reduces to a standard modeling framework with a single stage problem. When all decisions across multiple domains are negotiated jointly, intrahousehold bargaining power is represented by a single set of Pareto weights $\mu_j$ that apply uniformly across all domains (Chiappori & Mazzocco, 2017).

### 5.1.1 Efficiency

The traditional collective model of household decision-making typically rest on the assumption that allocations within the household are Pareto efficient (Chiappori, 1988, 1992). This assumption offers powerful analytical advantages. Most notably, it allows researchers to model and estimate household behavior without specifying the underlying bargaining process that drives decisions. Under efficiency, households obey decentralization properties akin to those in the fundamental welfare theorems, enabling substantial simplifications in the structure and estimation of household models.

While the framework presented here preserves the collective household framework's emphasis on joint decision-making, outcomes in our setting may not be Pareto efficient in the standard, unconditional sense. Instead, outcomes are *conditionally efficient*, meaning that the household reaches efficient allocations within each domain, given the outcomes of decisions in other domains. That is, in each domain $s$, the household maximizes a weighted sum of utilities subject to constraints shaped by past decisions $\phi^{s-1}$—including constraints on bargaining power, resource allocation, and feasible sets. In this regard, our framework mirrors the notion of conditional efficiency discussed in Lewbel & Pendakur (2022), where efficiency is defined relative to a fixed level of within-household cooperation. In our case, the fixed element is not cooperation per se, but the history of prior decisions across domains.

As discussed earlier, inefficiencies in the overall household allocation may arise from coordination costs, cognitive frictions, or other limitations on joint deliberation. These frictions render each decision effectively (partly or fully) irreversible, preventing the household from revisiting or renegotiating earlier choices even when doing so could lead to Pareto improvements. As a result, while decision-making in each domain may be efficient given the decisions taken on others, the household outcomes may be inefficient overall.

### 5.1.2 Extensions

While our simple framework takes the ordering of decision domains as exogenously given, the sequence in which issues are resolved may itself be an endogenous choice. The bargaining literature has emphasized that the order of decisions—commonly referred to as the *agenda*—can influence outcomes even when preferences and feasible sets are unchanged. Moreover, controlling the order

[20] While there is no explicit notion of time in our model, one could interpret this as the history of past decisions. Explicit dynamic considerations introduce a range of complexities, including uncertainty as well as the possibility of limited commitment to future agreements (Mazzocco, 2007). To keep the analysis tractable, we abstract from these challenges and develop a static framework in which decisions unfold sequentially but within a single time period (as in Lewbel & Pendakur (2022)).

in which decisions are made can be a source of power.[21]

Building on this insight, the framework can be extended to include an initial stage during which the household decides the order in which different areas of decision-making will be addressed. At one extreme, this sequencing may itself be the outcome of bargaining between spouses, reflecting their relative influence over agenda-setting. At the other extreme, the order may be determined unilaterally by one household member—for example, the husband. In many real-world settings, however, the sequence is shaped by external constraints, such as job opportunities, cultural norms, or unexpected shocks to income or health.

The framework can also be adapted to allow for joint deliberation over groups of issues, rather than assuming that all decisions are made one by one. In practice, some areas of household life are naturally considered together—for example, decisions about consumption and savings—while others, such as time use or health investments, may be addressed separately.[22] This more flexible structure captures the fact that some household decisions are made jointly, others sequentially, and many are shaped by both power dynamics and external constraints. It also implies that bargaining power can vary across domains, leading to outcomes that are efficient within each deliberation block but not necessarily across the household as a whole.

## 5.2 Discussion

The conceptual framework presented above, while retaining the core structure of the collective household model, explicitly highlights the possibility that intrahousehold bargaining power, as captured by the Pareto weights $\mu_j^s$, may vary across decision domains or issues. A household member who holds greater influence in decisions related to time allocation, for example, may not hold the same weight in decisions about consumption or health investment. This formulation moves away from the notion of a single measure of power and instead explicitly models intrahousehold power as a multi-dimensional construct, allowing for domain-specific asymmetries in decision-making authority and influence.

This theoretical foundation nicely maps onto our empirical analysis, which examines whether a unified measure of intrahousehold power is supported in the data, or whether different sets of predictors are needed to explain variation in each domain of intrahousehold relative wellbeing. Our results are consistent with the model's prediction that intrahousehold power is multidimensional and may vary across domains. We find that the survey questions most predictive of women's relative outcomes differ sharply across domains: questions that best predict time use outcomes are poor predictors of consumption or health, and vice versa.

There are several advantages to working within the collective household framework.[23] The ef-

[21] Multi-issue bargaining models highlight how agenda structure shapes outcomes through asymmetric information, strategic delay, and issue sequencing. For example, Bac & Raff (1996) and Busch & Horstmann (1999) show how strong players may delay hard issues to signal strength, while Busch & Horstmann (2002) illustrate how sequencing depends on ease or stakes of negotiation. Even under complete information, agenda order can affect outcomes (Fershtman, 1990; Lang & Rosenthal, 2001).

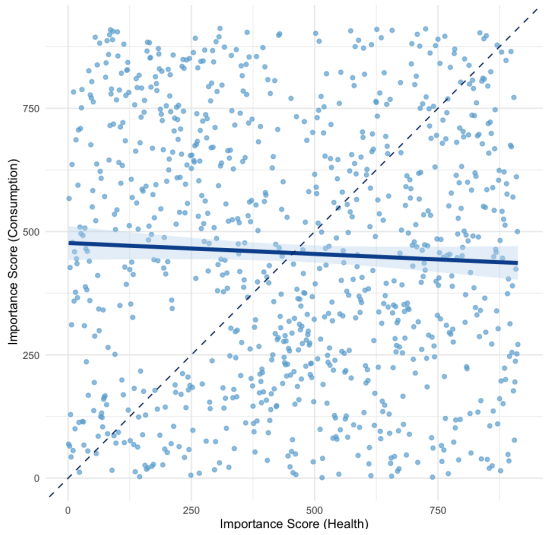
[22] Formally, this extension allows for decisions to be made in blocks of size $\kappa$, rather than one at a time or all at once. When $\kappa = 1$, the model reduces to a fully sequential structure. When $\kappa = S$, it corresponds to the standard collective household model with full joint optimization. Intermediate values of $\kappa$ capture settings where households can deliberate over multiple issues at once, but within practical limits.

[23] There exists alternative models of intrahousehold decision-making that can also rationalize the existence of domain-specific bargaining weights and the empirical patterns discussed in Section 4. For example, Lundberg & Pollak (1993) introduce the separate spheres bargaining model, where each spouse controls decisions within their respective domain—or "sphere" due to cultural norms. Other frameworks, including

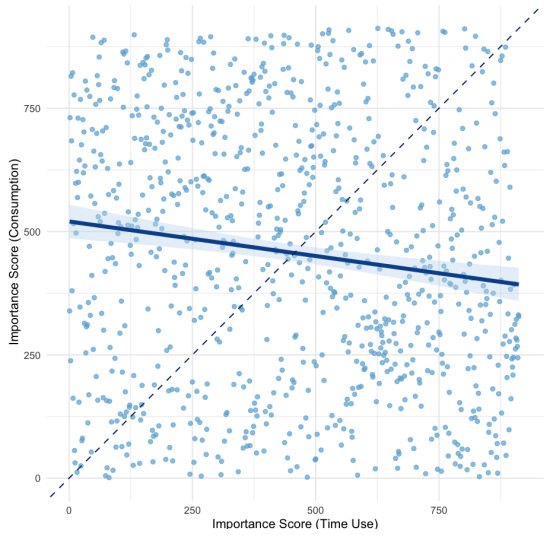
ficiency assumption at the core of the original collective model (Chiappori, 1988, 1992) plays a pivotal role in simplifying both theoretical formulation and empirical implementation. First, it allows researchers to abstract from the specifics of the bargaining process. Second, efficiency implies that household behavior satisfies decentralization principles, facilitating identification of key aspects of intrahousehold decision-making—such as bargaining power and the degree of joint consumption—when combined with additional structural assumptions (see, e.g., Lewbel & Pendakur, 2008; Dunbar *et al.*, 2013; Lewbel & Pendakur, 2022; Calvi *et al.*, 2023). Even when inefficiencies arise—as may be the case in our framework—Lewbel & Pendakur (2024) show that it is still possible to identify and estimate critical, otherwise unobserved features of the decision-making process. While the identification and the empirical estimation of our model are beyond the scope of this paper, extending efforts in that direction represents a promising avenue for future research.

issue-by-issue negotiation models with or without asymmetric information (as in Bac & Raff (1996); Busch & Horstmann (1999); Lang & Rosenthal (2001)), are also consistent with domain-specific bargaining weights.

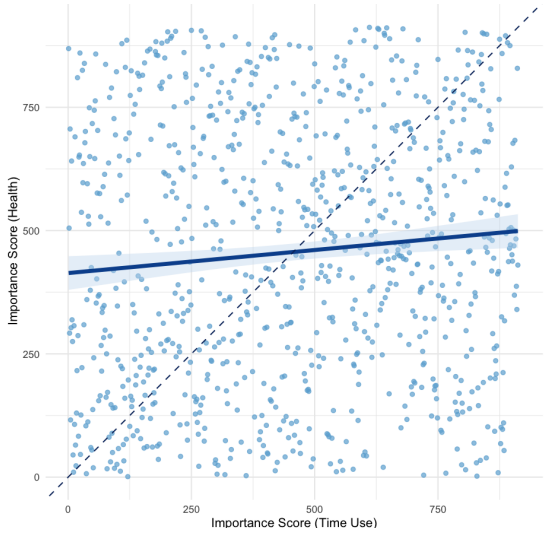
**Figure 1:** Correlation in Feature Importance of Power-related Survey Questions




(a) Consumption vs. Health



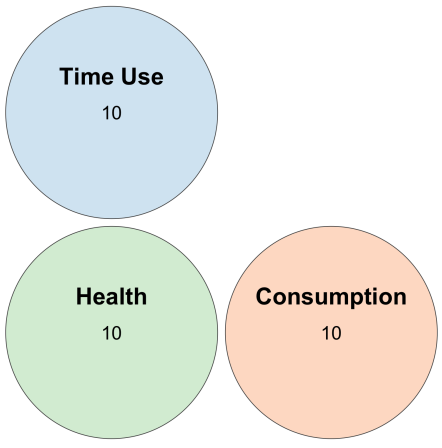

(b) Consumption vs. Time Use



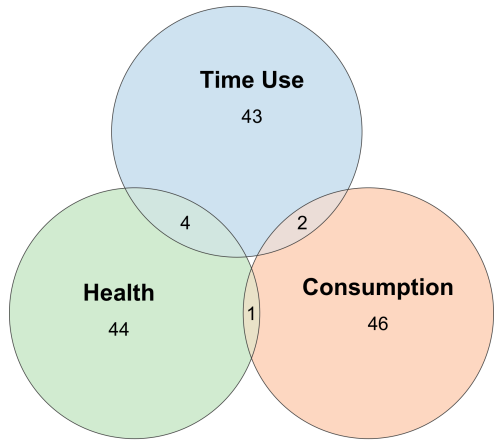

(c) Health vs. Time Use

NOTES: Relationship in the ranking of predictors across three domains of intrahousehold relative wellbeing using random forest permutation feature importance. Panel (a) compares consumption and health, panel (b) compares consumption and time use, and panel (c) compares health and time use. Each panel shows scatter plots with fitted lines. The 45-degree line indicates perfect agreement in rankings across domains.

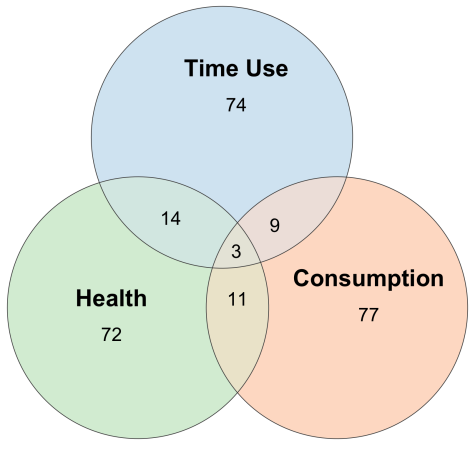
**Figure 2:** Overlap of Top Predictors Across Domains




(a) Top 10 predictors



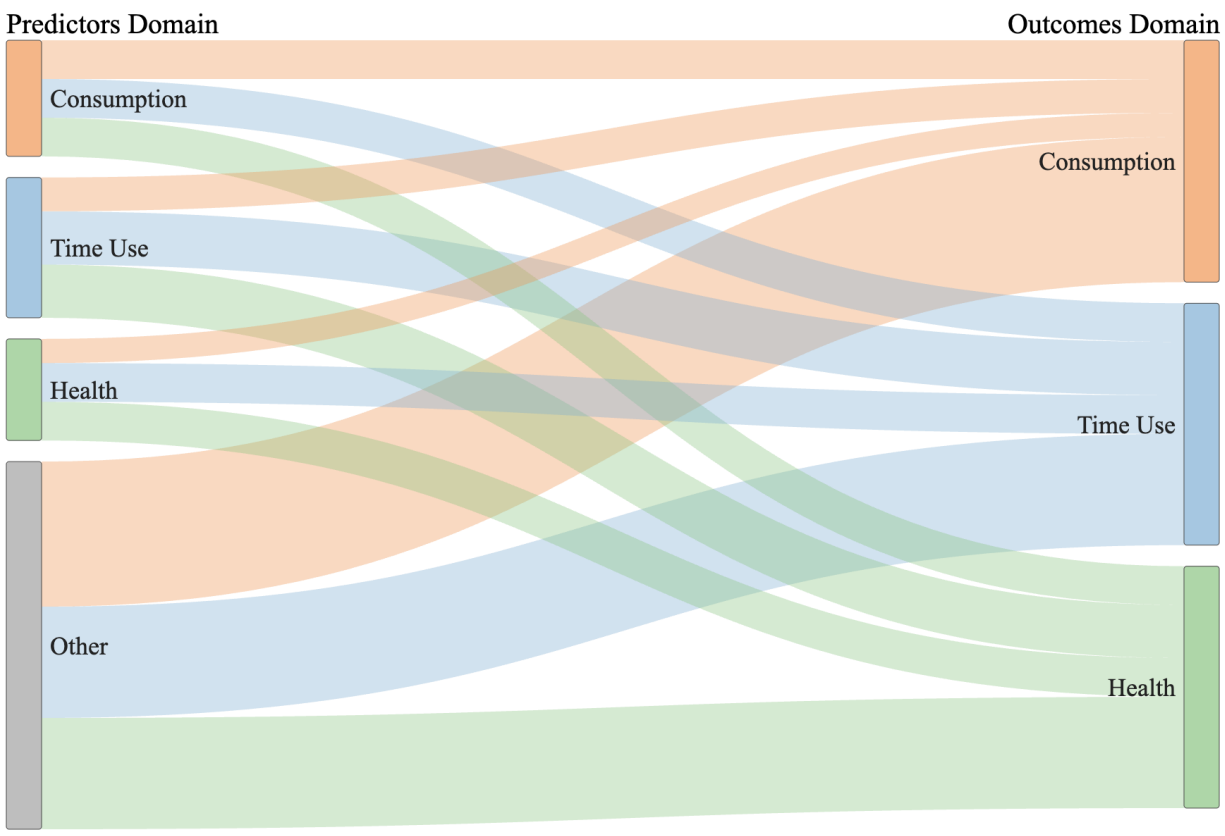

(b) Top 50 predictors




(c) Top 100 predictors

NOTES: Venn diagrams showing the overlap between top 10 (panel a), top 50 (panel b), and top 100 (panel c) predictors across the three domains of women's relative wellbeing based on the random forest permutation feature importance scores.

**Figure 3:** Multiple and Interconnected Dimensions of Intra-household Power




NOTES: The Sankey diagram illustrates the distribution of the top 50 survey questions (ranked by random forest permutation feature importance) across power domains and outcome dimensions. The width of each flow represents the share of variables from a given power domain that are among the top 50 predictors of women's relative wellbeing in consumption, time use, and health. The "Other" category includes power predictors related to fertility, religious expression, freedom of mobility, agricultural production (excluding time use), savings and loans, and non-farm asset decisions.

# References


Alkire, Sabina, Meinzen-Dick, Ruth, Peterman, Amber, Quisumbing, Agnes, Seymour, Greg, & Vaz, Ana. 2013. The women's empowerment in agriculture index. *World development*, **52**, 71–91. [2]

Almås, Ingvild, Armand, Alex, Attanasio, Orazio, & Carneiro, Pedro. 2018. Measuring and changing control: Women's empowerment and targeted transfers. *The Economic Journal*, **128**(612), F609–F639. [2]

Ambler, Kate, Doss, Cheryl, Kieran, Caitlin, & Passarelli, Simone. 2021. He says, she says: Spousal disagreement in survey measures of bargaining power. *Economic Development and Cultural Change*, **69**(2), 765–788. [6]

Anderson, Siwan. 2022. Unbundling female empowerment. *Canadian Journal of Economics/Revue canadienne d'économique*, **55**(4), 1671–1701. [4]

Argyriou, Andreas, Evgeniou, Theodoros, & Pontil, Massimiliano. 2006. Multi-task feature learning. *Advances in neural information processing systems*, **19**. [7], [22]

Bac, Mehmet, & Raff, Horst. 1996. Option contracts and renegotiation: A solution to the hold-up problem. *RAND Journal of Economics*, **27**(3), 554–572. [12], [13]

Basu, Kaushik. 2006. Gender and say: A model of household behaviour with endogenously determined balance of power. *The Economic Journal*, **116**(511), 558–580. [10]

Bayissa, Fitsum W, Smits, Jeroen, & Ruben, Ruerd. 2018. The multidimensional nature of women's empowerment: beyond the economic approach. *Journal of International Development*, **30**(4), 661–690. [8]

Breiman, Leo. 2001. Random forests. *Machine learning*, **45**, 5–32. [7], [21]

Brown, Caitlin, Calvi, Rossella, & Penglase, Jacob. 2021. Sharing the Pie: Undernutrition, Intrahousehold Allocation, and Poverty. *Journal of Public Economics)*. [5]

Browning, Martin, Chiappori, Pierre-Andre, & Lewbel, Arthur. 2013. Estimating Consumption Economies of Scale, Adult Equivalence Scales, and Household Bargaining Power. *The Review of Economic Studies*, **80**(4), 1267–1303. [2], [3], [9]

Busch, Lutz-Alexander, & Horstmann, Ignatius J. 1999. Incomplete contracts and the choice of dispute resolution. *Canadian Journal of Economics*, **32**(4), 956–977. [12], [13]

Busch, Lutz-Alexander, & Horstmann, Ignatius J. 2002. The game of negotiations: ordering issues and implementing agreements. *Games and Economic Behavior*, **41**(2), 169–191. [12]

Calvi, Rossella, Penglase, Jacob, Tommasi, Denni, & Wolf, Alexander. 2023. The more the poorer? Resource sharing and scale economies in large families. *Journal of Development Economics*, **160**, 102986. [13]

Campello, Ricardo JGB, Moulavi, Davoud, & Sander, Jörg. 2013. Density-based clustering based on hierarchical density estimates. *Pages 160–172 of: Pacific-Asia conference on knowledge discovery and data mining*. Springer. [26]

Chaturvedi, Sugat, Mahajan, Kanika, & Siddique, Zahra. 2024. Using Domain-Specific Word Embeddings to Examine the Demand for Skills. *Pages 171–223 of: Big Data Applications in Labor Economics, Part B*. Emerald Publishing Limited. [26]

Chiappori, Pierre-André. 1988. Rational Household Labor Supply. *Econometrica*, 63–90. [2], [3], [11], [13]

Chiappori, Pierre-André. 1992. Collective Labor Supply and Welfare. *Journal of Political Economy*, 437–467. [2], [3], [11], [13]

Chiappori, Pierre-Andre, & Mazzocco, Maurizio. 2017. Static and Intertemporal Household Decisions. *Journal of Economic Literature*, **55**(3), 985–1045. [11]

Donald, Aletheia, Koolwal, Gayatri, Annan, Jeannie, Falb, Kathryn, & Goldstein, Markus. 2020. Measuring women's agency. *Feminist Economics*, **26**(3), 200–226. [4]

Dunbar, Geoffrey R, Lewbel, Arthur, & Pendakur, Krishna. 2013. Children's Resources in Collective Households: Identification, Estimation, and an Application to Child Poverty in Malawi. *American Economic Review*, **103**(1), 438–471. [3], [13]

D'Souza, Anna, & Tandon, Sharad. 2019. Intrahousehold nutritional inequities in rural Bangladesh. *Economic Development and Cultural Change*, **67**(3), 625–657. [5]

Ewerling, Fernanda, Lynch, John W, Victora, Cesar G, van Eerdewijk, Anouka, Tyszler, Marcelo, & Barros, Aluisio JD. 2017. The SWPER index for women's empowerment in Africa: development and validation of an index based on survey data. *The Lancet Global Health*, **5**(9), e916–e923. [2]

Fershtman, Chaim. 1990. The importance of the agenda in bargaining. *Discussion Paper No. 689, Northwestern University, Center for Mathematical Studies in Economics and Management Science*. [12]

Ghuman, Sharon J, Lee, Helen J, & Smith, Herbert L. 2006. Measurement of women's autonomy according to women and their husbands: Results from five Asian countries. *Social Science Research*, **35**(1), 1–28. [6]

Jayachandra, Seema, & Voena, Alessandra. 2025. Women's power in the household. *Working Paper*. [4]

Jayachandran, Seema, Biradavolu, Monica, & Cooper, Jan. 2023. Using machine learning and qualitative interviews to design a five-question survey module for women's agency. *World Development*, **161**, 106076. [2], [4], [23]

Kabeer, Naila. 1999. Resources, agency, achievements: Reflections on the measurement of women's empowerment. *Development and change*, **30**(3), 435–464. [4]

Lang, Kevin, & Rosenthal, Robert W. 2001. Bargaining piecemeal or all at once? *The Economic Journal*, **111**(473), 526–540. [12], [13]

Laszlo, Sonia, Grantham, Kate, Oskay, Ecem, & Zhang, Tingting. 2020. Grappling with the challenges of measuring women's economic empowerment in intrahousehold settings. *World Development*, **132**, 104959. [4]

Lewbel, Arthur, & Pendakur, Krishna. 2008. Estimation of collective household models with Engel curves. *Journal of Econometrics*, **147**(2), 350–358. [13]

Lewbel, Arthur, & Pendakur, Krishna. 2022. Inefficient collective households: Cooperation and consumption. *The Economic Journal*, **132**(645), 1882–1893. [3], [11], [13]

Lewbel, Arthur, & Pendakur, Krishna. 2024. Estimating a model of inefficient cooperation and consumption in collective households. *Review of Economics of the Household*, **22**(3), 865–907. [13]

Lundberg, Shelly, & Pollak, Robert A. 1993. Separate spheres bargaining and the marriage market. *Journal of political Economy*, **101**(6), 988–1010. [4], [12]

Manski, Charles F, & Molinari, Francesca. 2008. Skip sequencing: A decision problem in questionnaire design. *The annals of applied statistics*, **2**(1), 264. [6]

Mazzocco, Maurizio. 2007. Household intertemporal behaviour: A collective characterization and a test of commitment. *The Review of Economic Studies*, **74**(3), 857–895. [11]

McInnes, Leland, Healy, John, & Melville, James. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. [26]

Meinshausen, Nicolai, & Bühlmann, Peter. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473. [7], [23]

Quisumbing, Agnes R. 2025. From bargaining power to empowerment: Measuring the unmeasurable. *Agricultural Economics*, **56**(3), 419–430. [2]

Reimers, Nils, & Gurevych, Iryna. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Pages 3982–3992 of: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. [25]

Sen, Amartya. 1990. Development as capability expansion. *The community development reader*, **41**, 58. [4]

Sen, Amartya, *et al.* 1999. Commodities and capabilities. *OUP Catalogue*. [4]

Song, Kaitao, Tan, Xu, Qin, Tao, Lu, Jianfeng, & Liu, Tie-Yan. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, **33**, 16857–16867. [25]

Strobl, Carolin, Boulesteix, Anne-Laure, Zeileis, Achim, & Hothorn, Torsten. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, **8**(1), 1–21. [22]

Yuan, Ming, & Lin, Yi. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68**(1), 49–67. [7], [23]

# Supplemental Appendix

## Contents

# A Methodology: Details

In this section, we describe our approach to select survey questions that are the most predictive of women's relative outcomes within a household. We perform this analysis separately for questions asked only to women, only to men, and for the combined men's and women's questionnaires. Additionally, we perform the analysis separately for each welfare dimension. Since female empowerment is inherently multifaceted, even within a single welfare dimension, we frame the selection of survey questions as a multi-task problem, as detailed below.

## A.1 Random Forest

Our preferred approach uses the random forest regressor which supports multiple outcome variables and often achieves a high degree of predictive accuracy. It is an ensemble method that fits a large number of decision trees on bootstrap samples (drawn with replacement) from the training set. Essentially, decision trees recursively partition the data using a complex set of if-else decision rules based on cut-off values of features. The cut-off values are chosen to group together the data into subsets (or nodes) such that the variance in the outcome value within each subset is minimized. This ensures that observations with similar output values belong to the same partition. The outcome in the final subsets of the data (known as the leaf nodes) is predicted by taking the average of outcome in the training data belonging to each node. Note that individual decision trees tend to overfit the training data. Therefore, random forest injects two sources of randomness. First, as discussed above, it builds each tree using a bootstrap sample. In addition, each node in individual decision trees is split based on a random subset of features. The randomness from these two sources results in a low correlation between prediction errors of individual decision trees. These predictions are then averaged to offset the errors which prevents overfitting the model on the training data. We run the random forest regressor 100 times on 50 percent random subsets of the data and use 5-folds cross validation within each iteration to choose the best hyperparameters.[24] The remaining 50 percent observations in each iteration are set aside and comprise the out-of-bag sample.

To compute feature importance, we use permutation feature importance introduced by Breiman (2001). For a given feature, this is calculated by measuring the decrease in model score (in this case, $R^2$) after randomly shuffling that feature. We follow the steps below for every iteration of random forest:

1. Fit the model on 50 percent subsample of the data and, on the remaining 50 percent out-of-bag data denoted as $(X^{oob}, Y^{oob})$, compute $R^2$ (or $score^{oob}$) for the random forest regressor:

2. For each feature $x_j \in X$, repeat the following L times (we use L=10):

    - Randomly shuffle (or permute) the feature to create a modified version of the data $(\tilde{X}^{oob}, Y^{oob})$

[24] The two most important hyperparameters for random forest are the number of trees and the number of random subset of features considered for splitting a node. A larger number of trees improves predictive accuracy at a decreasing rate while increasing computational time. Therefore, we set the number of trees to 1,000 guided by the trade-off between predictive accuracy and computational time. For our second hyperparameter, i.e. the fraction of features to consider when looking for the best split we use the hyperparameter values $[.05, .075, .1, .125, .15]$, and choose the value that gives the best cross-validation score.

- On this modified data and for every repetition $l \in \{1, 2, ..., L\}$ compute $R^2$ (or $score^{oob}_{j,l}$)

3. The feature importance of variable $x_j$ is given as $F_j = score^{oob} - 1/L \sum_{l=1}^{L} score^{oob}_{j,l}$

We then compute the mean permutation feature importance $\bar{F}_j$ over 100 random forest iterations for each feature. Therefore, permutation feature importance can be interpreted as the average decrease in $R^2$ on out-of-bag sample when the feature is shuffled. This method overcomes two shortcoming of mean decrease in impurity (MDI)—a commonly used alternative measure for computing feature importance.[25] First, impurity-based feature importance methods inflate the importance of continuous features relative to categorical ones (Strobl *et al.*, 2007). This is because features with high cardinality have more split points, and therefore, are more likely to reduce variance in the child node by chance alone. This is important for us since our predictors include a combination of categorical and continuous variables. Secondly, permutation feature importance can be computed over the out-of-bag sample. This makes it less sensitive to giving high importance to features that are not predictive on the out-of-bag sample when there is overfitting on the training data. Finally, in the interest of maintaining consistency with the alternative methods discussed below and to take into account correlation between features, we rank features based on $\bar{F}_j$ and plot the adjusted $R^2$ by running a linear regression for each of the outcome measures against the top k predictors.

## A.2 Multi-task Lasso Stability Selection

Our second approach uses multi-task Lasso proposed by Argyriou *et al.* (2006), which is an extension of the linear Lasso model but allows us to simultaneously select predictors for multiple outcomes. It jointly selects a sparse set out of J features for multiple (K) regression problems (or tasks) by constraining the set of selected features to be the same across tasks. Consider the linear model:

$$\hat{Y} = X\beta$$

where X represents the set of features (or predictors), Y is the matrix of target variable (or outcomes), and $\beta$ represents the coefficients. To constrain the set of selected features to be the same across tasks, multi-task Lasso uses a mixed $l_1 l_2$ norm for regularization and solves the following optimization problem:

$$\min_{\beta} ||Y - X\beta||^2_{Fro} + \frac{1}{\lambda}||\beta||_{21}$$

where $\lambda$ is a constant known as the regularization parameter and Fro indicates the Frobenius norm:

$$||Y - X\beta||_{Fro} = \sqrt{\sum_{k=1}^{K}\sum_{i=1}^{N}(y_{i,k} - x_i\beta_k)^2}$$

where $i$ indexes an observation, $k$ indexes the task, and $j$ indexes the parameter. In our case, the outcomes are different measures of relative welfare within a dimension. The $l_1 l_2$ mixed norm is given

[25] MDI computes feature importance by measuring the sum of decrease in within-node variance of outcome (or impurity) over all nodes split using that feature, averaged over all trees in the forest. In other words, it is the difference in the impurity of a node and the weighted (by fraction of samples reaching that node) sum of impurity of child nodes obtained by splits on a feature.

by:

$$||\beta||_{21} = \sum_{j=1}^{J} \sqrt{\sum_{k=1}^{K} \beta_{j,k}^2}$$

We combine this with the stability selection algorithm wherein we repeatedly run multi-task Lasso on sub-samples of our data. Stability Selection was proposed by Meinshausen & Bühlmann (2010) and has previously been applied by Jayachandran *et al.* (2023) using Lasso. We instead use multi-task Lasso and run the algorithm separately for the different welfare dimensions using the following steps. First, we take a 50% sample of observations without replacement. We then run multi-task Lasso with 5-folds cross-validation to choose the regularization parameter.[26] This gives us a sparse set of selected predictors, i.e. for which the coefficients do not shrink to 0. We then repeat this process for 1,000 iterations and order the predictors based on the number of times they are selected for each dimension. Finally, we regress each of the outcome measures on the top k predictors and record the adjusted $R^2$ obtained. We also check the sensitivity of our results using Lasso stability selection, which does not impose the restriction of selecting the same set of features across different outcomes within a dimension.

## A.3 Group Lasso Stability Selection

As described in Section 2.2, we re-code responses to questions with categorical answers into multiple binary variables. Therefore, to identify which questions (as opposed to which variables) are predictive of women's empowerment, our second approach uses group lasso (Yuan & Lin, 2006). To implement this, we group together indicator variables created from a single (or merged) categorical question while keeping questions with continuous/ranked responses and binary responses as part of distinct groups. In other words, each unique survey question is mapped to a group of predictors resulting in G groups. We solve a similar optimization problem as before:

$$\min_{\beta} ||Y - \sum_{g=1}^{G} [X_g \beta_g]||_{Fro}^2 + \frac{1}{\lambda}||\beta||_{21}$$

For notational convenience, we separately include coefficients for group g in the Frobenius norm:

$$||Y - \sum_{g=1}^{G} [X_g \beta_g]||_{Fro} = \sqrt{\sum_{k=1}^{K} \sum_{i=1}^{N} (y_{i,k} - \sum_{g=1}^{G} [x_{i,g} \beta_{g,k}])^2}$$

However, the penalty or regularization term is slightly modified to impose group-wise sparsity:

$$||\beta||_{21} = \sum_{g=1}^{G} \sqrt{\rho_g} \sqrt{\sum_{k=1}^{K} \sum_{j=1}^{\rho_g} \beta_{j,g,k}^2}$$

[26] For pre-processing, we remove duplicate (or perfectly correlated) predictors and features with very little variation (i.e. those that vary less than 5 times in our data). We also remove outliers—observations for which relative calorie consumption is 0 or 1. To ensure that the continuous variables do not dominate the objective function, we scale them to a range between 0 and 1 before using the coordinate descent algorithm to fit the model.

where $\rho_g$ is the number of predictors in group g such that $\sum_{g=1}^{G} \rho_g = J$. This penalty term jointly selects a sparse set of groups and ensures that all the predictors created out of a single categorical variable are included and excluded jointly, while restricting the set of groups to be the same across tasks. Note that if each covariate belongs to its own group, then this would reduce to multi-task lasso. We follow the same pre-processing steps and combine group lasso with the stability selection procedure described in the previous sub-section, and rank variables based on their selection frequency.

# B Additional Results: Details

## B.1 Target Respondents

Figure A4 in the Appendix reports the share of the most predictive survey questions that originate from the women's questionnaire relative to the men's questionnaire, for varying thresholds of top predictors (Top 10, Top 20, and so on). Across domains, women's survey questions account for a larger share of the most predictive variables, indicating that women's self-reported responses are, on average, more informative for explaining relative well-being outcomes than men's responses.

Importantly, however, men's responses are not redundant. Although women's survey questions dominate the top predictor lists, adding men's questions increases explanatory power. For instance, adjusted $R^2$ is higher when the predictive models are estimated using the combined questionnaire rather than either men's or women's questions alone. This suggests that men and women provide complementary information about intra-household dynamics, even if women's responses tend to carry more predictive weight. From a survey design perspective, these patterns imply that when only one respondent can be interviewed, surveying women may yield stronger predictive content for key well-being outcomes. However, when feasible, collecting information from both spouses is valuable: men's responses contribute additional predictive power, capturing dimensions of intra-household behavior not fully reflected in women's responses alone.

## B.2 Predictive Power

A natural question is how accurately survey-based measures of intra-household power predict relative spousal outcomes. From a policy perspective, strengthening women's power is often presumed to improve their relative well-being, so it is important to assess how well standard survey questions forecast outcomes across domains. Figure A5 in the Appendix plots adjusted $R^2$ against the top $k$ predictors selected by a random forest, ordered by feature importance. Each curve considers up to the 50 most important predictors, where each predictor is a survey question interacted with respondent gender. The level of adjusted $R^2$ reflects overall predictive power, while the shape of the curve shows how quickly it accumulates as additional predictors are included.

Time use outcomes (Panel A) are the most predictable: market and domestic work measures reach adjusted $R^2$ values near 0.15–0.20, with rapid gains from the top 20 predictors. Health outcomes (Panel B) are more uneven: the domestic violence index exhibits strong predictive power, while other

health measures remain difficult to predict. Consumption outcomes (Panel C) display moderate, steady gains, but explanatory power remains below 0.10 even with many predictors. These patterns suggest that a handful of survey items capture meaningful variation in certain domains, while others contain little predictive information. The curves are generally concave, indicating that most of the predictive power is concentrated in a relatively small set of questions. Importantly, these figures represent sizable gains over predictive models based on demographic variables alone (Table A2, Column 5).[27]

# C Classifying Questions into Power Domains using Sentence Transformer

In the main analysis, we rely on a manual categorization of survey questions into *power domains*. To assess the validity of this approach, we also cluster the questions into domains based on text embeddings obtained using a sentence transformer model and compare the classification with the manual categories. The procedure involves the below four steps:

1. **Question-domain pairs.** We first use the BIHS questionnaire to construct a list of questions (e.g., "Did you participate in [ACTIVITY] in the past 12 months? How much input did you have in making decisions about [ACTIVITY]?") along with the domain to which each variable refers (e.g., "Fishing or fish culture"). We then remove duplicates to retain 147 unique question-domain combinations, so that questions with more response categories do not receive a disproportionate weight during clustering.

2. **Embedding generation.** We generate embeddings (vector representations) for each question ($v_q$) and its associated domain ($v_g$) using a `sentence transformer` model (Reimers & Gurevych, 2019). Conventional transformers such as BERT or RoBERTa are less suitable for this task, as they obtain sentence representations by averaging contextualized words embeddings. This limits their effectiveness for applications that depend on semantic similarity. The sentence-transformer framework addresses this by fine-tuning transformer models to produce sentence-level embeddings that capture the overall semantic meaning. Specifically, we use the pre-trained `all-mpnet-base-v2` model, a fine-tuned variant of Microsoft's *mpnet-base* (Song *et al.*, 2020), trained on over a billion sentence pairs from diverse sources (e.g., academic papers, Wikipedia, Reddit, and Stack Exchange). This model encodes text into a 768-dimensional vector space and ranks among the top-performing models for *semantic textual similarity*.

   Given a set of $N$ sentence pairs $(v_i^1, v_i^2)$, the model computes embeddings $\vec{v}_i^1$ and $\vec{v}_i^2$ for $i \in \{1, ..., N\}$ by taking the mean of contextualized word embeddings using a conventional transformer. It then calculates the cosine similarity $cos(\vec{v}_i^1, \vec{v}_j^2)$ for all possible sentence pairs, forming a similarity matrix $M$ whose elements $M_{i,j}$ represent the pairwise cosine similarity between sentences $v_i^1$ and $v_j^2$. Finally, the embeddings are fine-tuned by minimizing a symmetric cross-entropy loss that maximizes the similarity between true sentence pairs relative to all others:

[27] One concern is that individual categorical responses may be weak predictors, with explanatory power arising from patterns across items. We therefore estimate group Lasso models that select variables at the survey-question level; the main results are unchanged.

$$l = -\frac{1}{2N}\sum_{i=1}^{M}\left( log\frac{exp(cos(\vec{v}_i^1, \vec{v}_i^2))}{\sum_{j=1}^{N} exp(cos(\vec{v}_i^1, \vec{v}_j^2))} + log\frac{cos(\vec{v}_i^2, \vec{v}_i^1)}{\sum_{j=1}^{N} exp(cos(\vec{v}_i^2, \vec{v}_j^1))}\right)$$

In our context, this ensures that conceptually related questions—such as those reflecting similar aspects of bargaining power—are located closer together in the embedding space and form smaller angles between their corresponding vectors, while dissimilar questions are positioned further apart. Semantic similarity can then be quantified using cosine similarity. Furthermore, sentence transformers are highly computationally efficient, operating orders of magnitude faster than conventional transformer models.

3. **Composite embedding.** For each question, we construct a composite embedding

$$v = (1-w)v_q + wv_g,$$

where $w$ denotes the relative weight placed on the domain. In the baseline specification, we set $w = 0.6$, reflecting the idea that domains should play a more central role in classification while still retaining information from the questions. If the domain is missing, we set $w = 0$. This choice yields results most consistent with our manual categorization (as discussed below). For comparison, we also consider $w = 0.9$, which appears to perform slightly better on visual inspection. Finally, we systematically vary $w$ from 0 to 1 in increments of 0.05 to examine the sensitivity of the results to the relative importance placed on domains versus questions.

4. **Clustering.** We first reduce the 768-dimensional embeddings to 15 dimensions using Uniform Manifold Approximation and Projection (UMAP)—a non-linear dimensionality reduction method that preserves the underlying topological structure of the data (McInnes *et al.*, 2018).[28] This dimension reduction step improves the effectiveness of clustering algorithms, which tend to perform poorly in very high-dimensional spaces.[29] We then apply the k-means clustering algorithm to partition the survey questions into 10 domains with the objective of minimizing the within-cluster sum of squared Euclidean distances (inertia). The algorithm begins by randomly initializing cluster centroids and then proceeds iteratively: each point is assigned to the nearest centroid, and the centroids are updated as the mean of the points in each cluster. This process continues until assignments stabilize.[30] We obtain similar results using HDBSCAN, a hierarchical, density-based clustering algorithm (Campello *et al.*, 2013).[31]

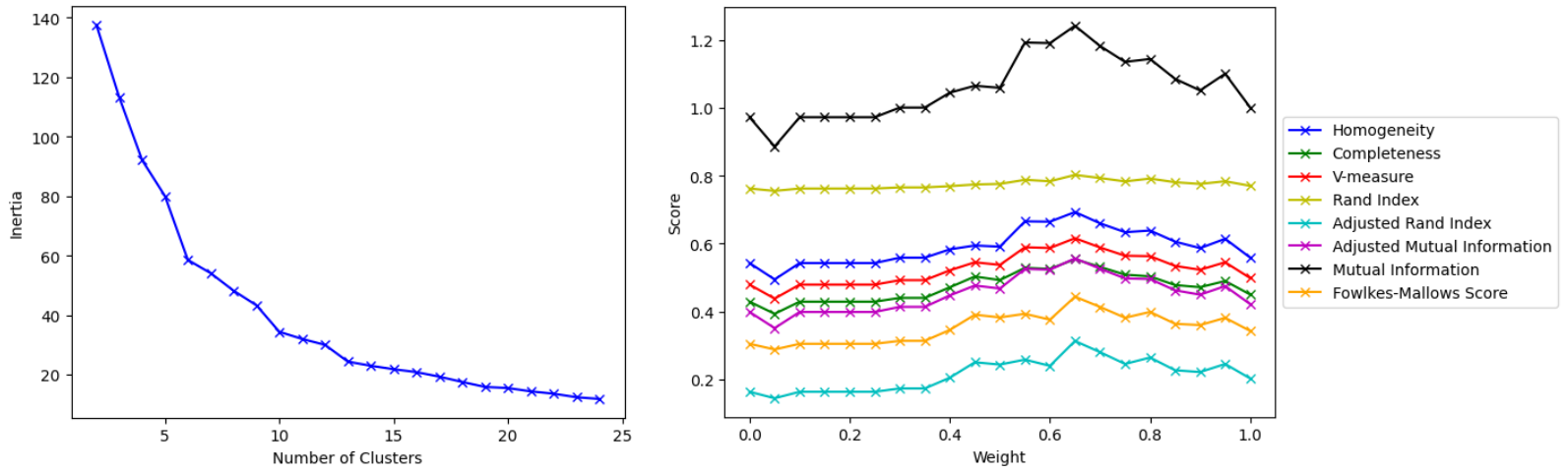
[28]The key hyperparameter in UMAP is the number of nearest neighbors. Since our dataset contains 147 questions, each question can have at most 146 neighbors. We therefore set this parameter to 146 to capture the global structure of the data. We also set the minimum distance between points to 0, which produces a denser low-dimensional representation more suitable for clustering, and use cosine distance as the similarity metric.

[29]This "curse of dimensionality" arises because Euclidean distances lose their discriminatory power as dimensionality increases: the gap between the nearest and farthest points shrinks, so observations appear almost equally distant. At the same time, the data become sparse, leaving large empty regions that make it difficult for algorithms to detect dense, well-defined clusters.

[30]To determine the number of clusters, we use the elbow method by plotting inertia as the number of clusters increases from 2 to 25 and selecting the point where the curve flattens, indicating diminishing returns from additional clusters. We depict this in Figure A1, Panel (a).

[31]For HDBSCAN, we set the *minimum cluster size* to 4 (ensuring that each domain contains at least four questions), *minimum samples* to 1 (so that even a single point in a dense region is treated as a core point and fewer questions are discarded as "noise"), and $\epsilon = 0.65$, so that clusters closer than this threshold in Euclidean distance within the UMAP space are not split further. For an accessible introduction to these methods and their application in economics, see Chaturvedi *et al.* (2024), who use them to analyze online job ad texts.

**Figure A1:** Domain Classification




(a) Inertia vs. # of clusters ($w = 0.6$)

(b) Cluster evaluation metrics

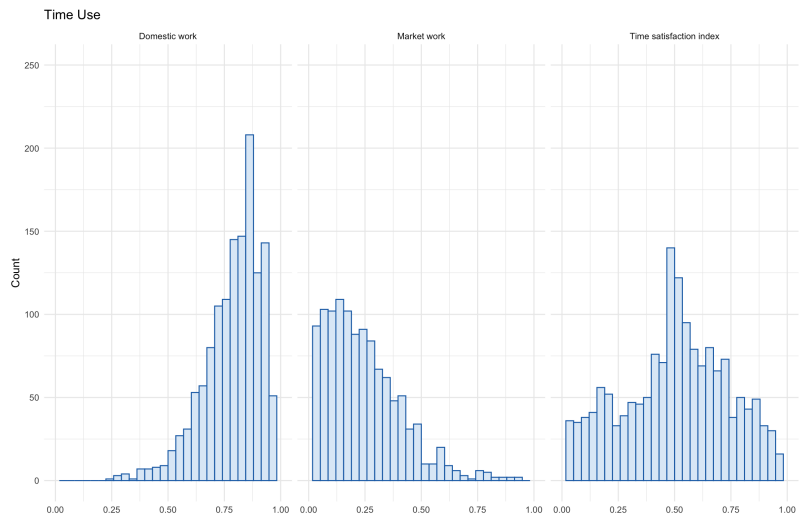
Note: Panel (a) shows the elbow method, where the optimal number of clusters is chosen at the point where inertia flattens. Panel (b) compares the clusters against the manually constructed domains, using standard metrics including homogeneity, completeness, V-measure, Rand index, adjusted Rand index, adjusted and unadjusted mutual information, and Fowlkes–Mallows score, for 10 clusters across different values of $w$.

We compare the clusters against the manually constructed domains using standard measures, including homogeneity, completeness, V-measure, Rand index, adjusted Rand index, mutual information (adjusted and unadjusted), and the Fowlkes–Mallows score. The results are shown in Figure A1, Panel (b) for 10 clusters for different values of $w$. Across specifications, the clustering results align closely with the manual classification, suggesting that the findings reported in the main analysis are not sensitive to the particular classification method employed.[32] We find that the concordance with the manual categorization tends to be the highest for weight $w$ around 0.6–0.7. For example, with $w = 0.6$, the homogeneity and completeness scores are 0.66 and 0.53, respectively. In practice, this implies that about 66% of the questions within a cluster belong to the same manually classified domain, while 53% of the questions in a given manual domain are grouped into the same cluster.

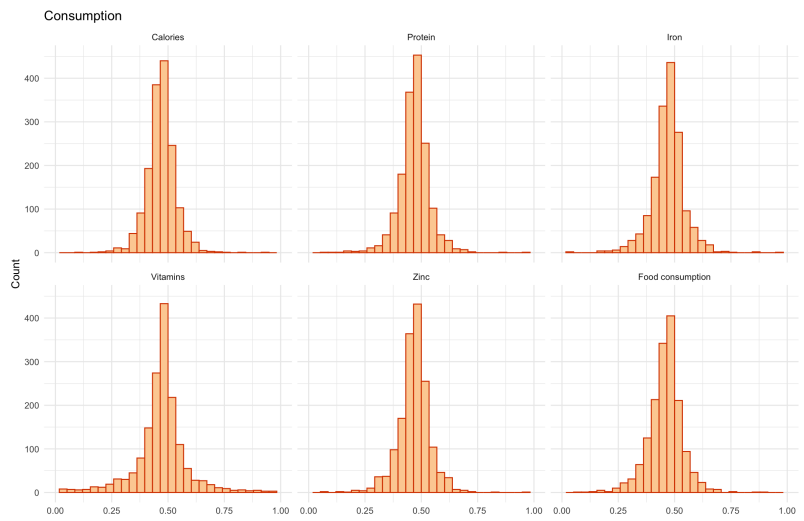
[32] For detailed definitions of these metrics, see https://scikit-learn.org/stable/modules/clustering.html.

# D Additional Figures

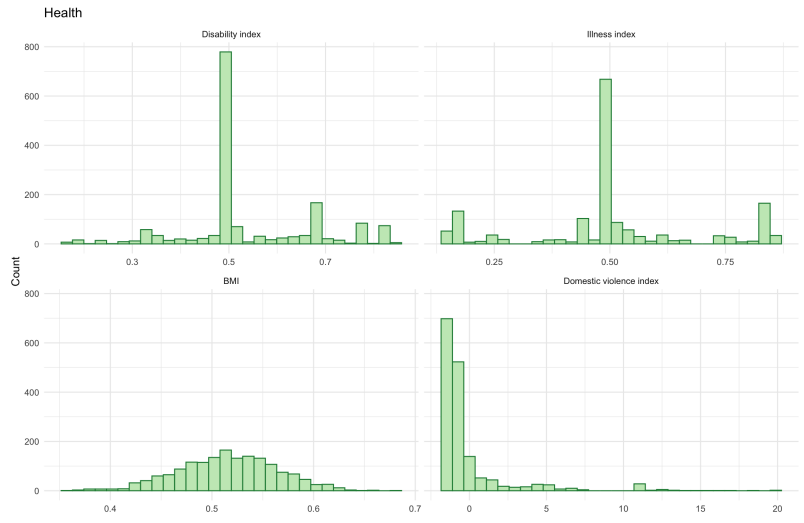
**Figure A2:** Intra-household Relative Outcomes in Various Domains




(a) Time use




(b) Consumption




(c) Health

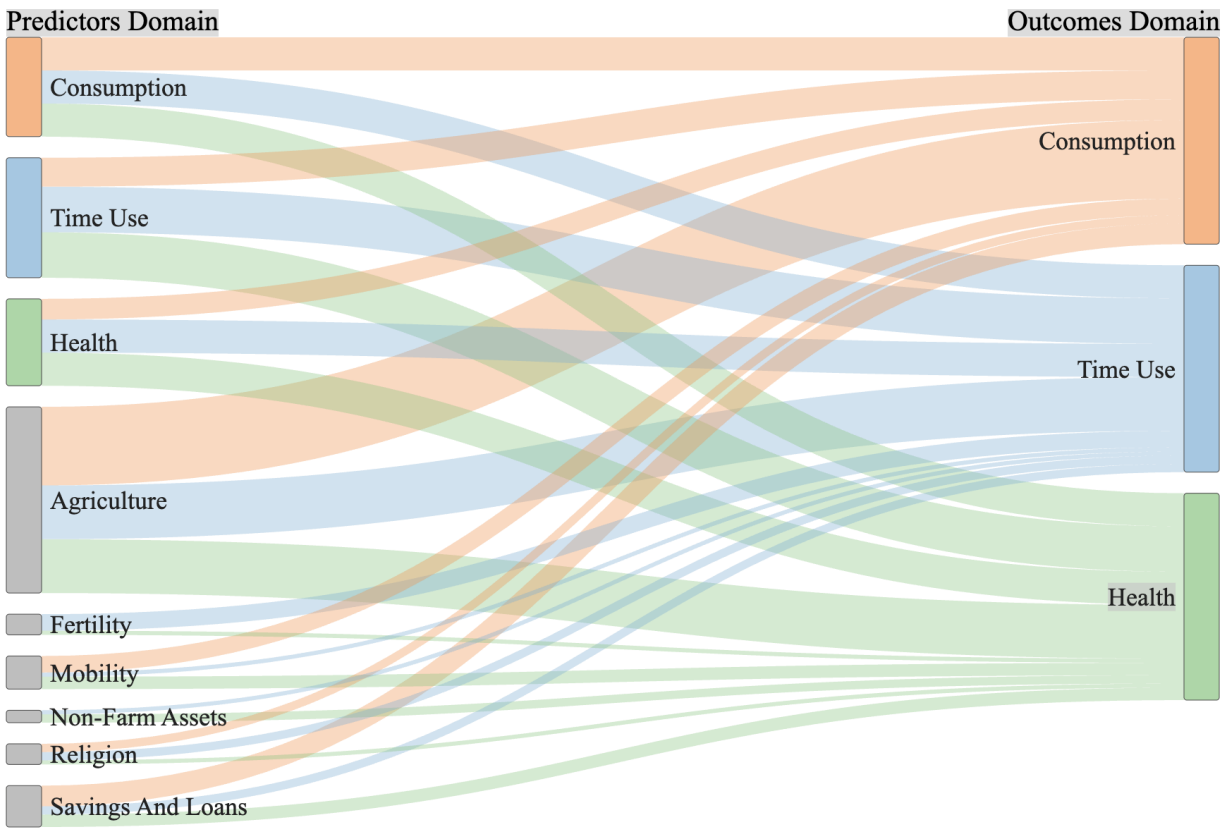
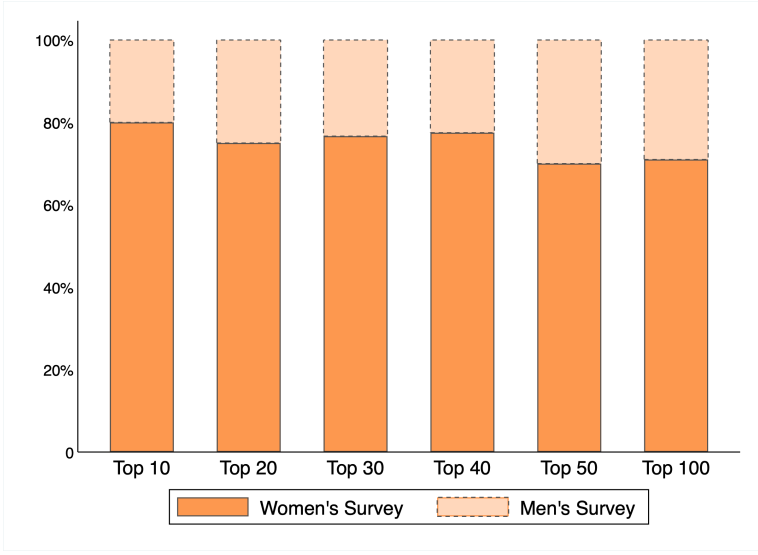


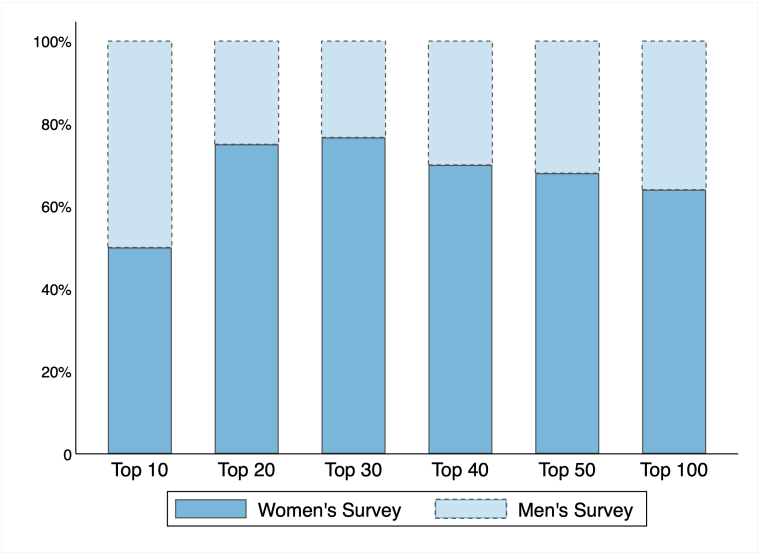

**Figure A3:** Multiple Dimensions of Intra-household Power (Details)

NOTES: The Sankey diagram illustrates the distribution of the top 50 survey questions (ranked by random forest permutation feature importance) across power domains and outcome dimensions. The width of each flow represents the share of variables from a given power domain that are among the top 50 predictors of women's relative well-being in consumption, time use, and health.

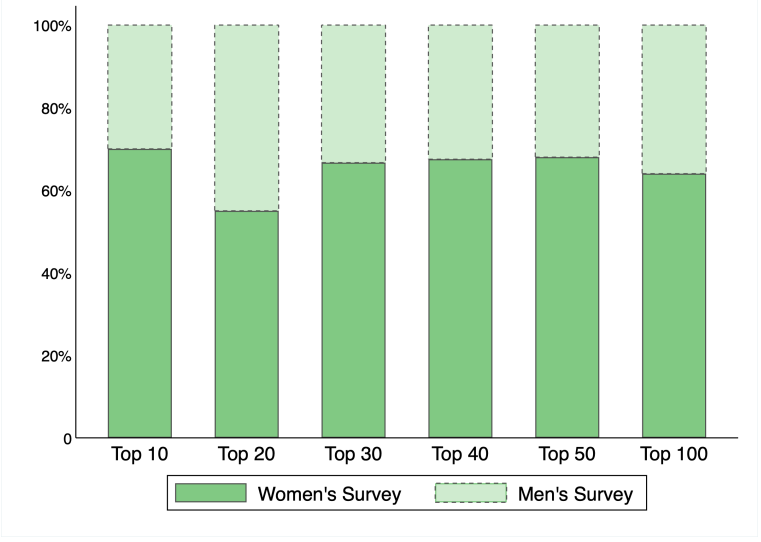
**Figure A4:** Share of Top Predictors by Respondent Gender




(A) Consumption



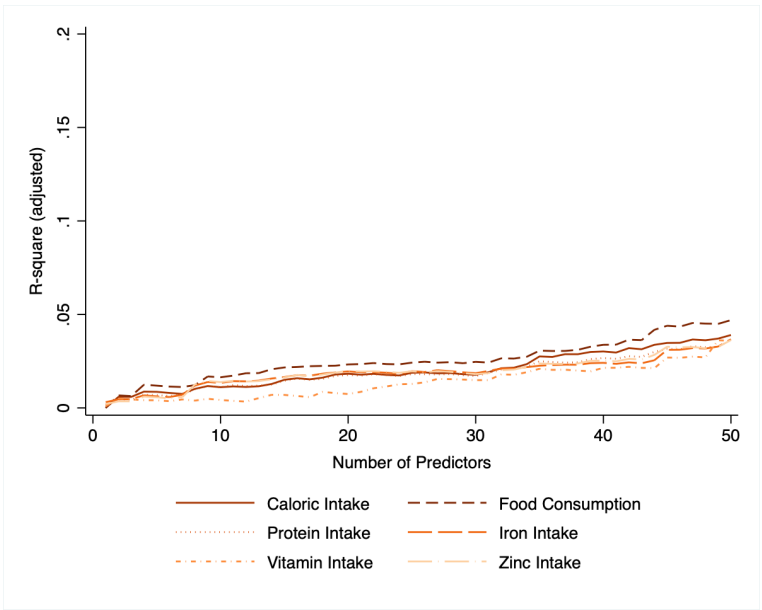

(B) Time Use



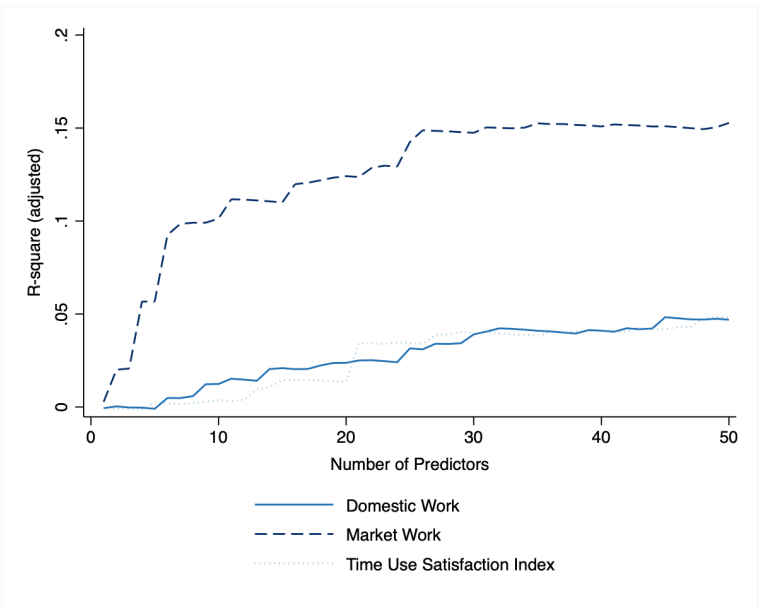

(C) Health

NOTES: Random Forest Estimation. The dark bars represent the share of the top X questions ranked by feature importance that are asked to women, with the light bars representing the share asked of men.

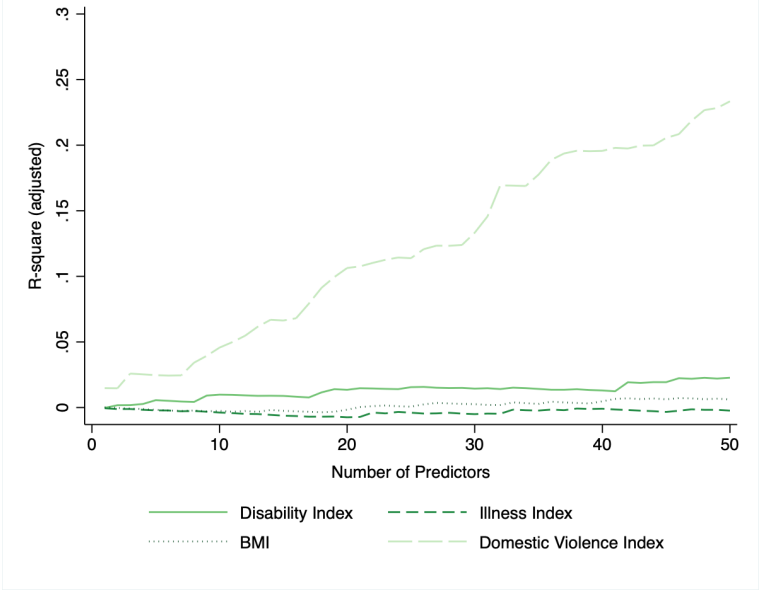
**Figure A5:** Predictive Power of Power-related Questions




(A) Consumption




(B) Time Use




(C) Health

NOTES: This figure shows the predictive performance of survey-based measures of intra-household power for different well-being outcomes. For each domain—time use, consumption, and health—we plot the adjusted $R^2$ obtained from linear regressions as we incrementally add the top $k$ predictors, ranked by random forest feature importance.

# E Additional Tables

**Table A1:** Module Overview

| Module | Description | Questions for Women | Questions for Men |
|---|---|---|---|
| *Women's Status* | | | |
| Z1 | Work Earnings and Expense | 11 | - |
| Z2 | Freedom of Mobility | 10 | - |
| *Women's Empowerment in Agriculture Index (WEAI)* | | | |
| WEA | Ability to Be Interviewed Alone | 1 | 1 |
| WE2 | Role in Household Decision-Making Around Production and Income Generation | 16 | 16 |
| WE3A | Access to Productive Capital | 54 | 48 |
| WE3D | Access to Loans | 21 | 25 |
| WE5A | Decision Making in the Household | 21 | 15 |
| WE5B | Decision Making Vignettes | 12 | 12 |
| WE5C | Motivation for Decision Making | 32 | 32 |
| WE6B | Satisfaction with Time Allocation | 2 | 2 |
| Total | | 172 | 143 |

Note: Only survey questions used in the estimation are included in the table.

**Table A2:** Target Variables

| | Descriptive Statistics | | | | Explanatory Power of Demographics |
|---|---|---|---|---|---|
| | Mean | Median | Std. Dev | Obs. | $R^2$ |
| *Time Use* | | | | | |
| Domestic Work | 0.822 | 0.844 | 0.143 | 1620 | 0.023 |
| Market Work | 0.208 | 0.151 | 0.232 | 1620 | 0.075 |
| Time Use Satisfaction | 0.514 | 0.508 | 0.226 | 1620 | 0.055 |
| *Food Consumption* | | | | | |
| Food Consumption | 0.461 | 0.465 | 0.080 | 1620 | 0.024 |
| Caloric Intake | 0.471 | 0.471 | 0.069 | 1620 | 0.019 |
| Protein Intake | 0.471 | 0.472 | 0.073 | 1620 | 0.018 |
| Iron Intake | 0.472 | 0.474 | 0.078 | 1620 | 0.019 |
| Vitamin Intake | 0.472 | 0.477 | 0.118 | 1620 | 0.019 |
| Zinc Intake | 0.470 | 0.472 | 0.076 | 1620 | 0.017 |
| *Health* | | | | | |
| Health Index | 0.540 | 0.500 | 0.133 | 1620 | 0.020 |
| Illness Index | 0.505 | 0.500 | 0.185 | 1620 | 0.017 |
| BMI | 0.516 | 0.516 | 0.049 | 1620 | 0.037 |
| Domestic Violence Index | 0.000 | -0.612 | 2.599 | 1620 | 0.014 |

Note: BIHS data. Variables are ratios computed as $\frac{y_w}{y_w+y_h}$ where $y_w$ and $y_h$ denote the values of some variable for the wife and husband, respectively, with the exception of the domestic violence index which is only computed for women. Time Use Satisfaction is computed using a principal components analysis of several survey questions on stated satisfaction. The $R^2$ refers to a regression of the outcome variable on standard demographic information of the both spouses including their education, age, age$^2$, literacy, and the number of children in the household.

**Table A3:** Random Forest Food Consumption Variables

| Question | Answer | Feature Importance |
|---|---|---|
| Your actions with respect to how to express religious faith are motivated by a desire to avoid blame or so that other people speak well of you? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Somewhat true | 0.0006 |
| Did you participate in raising poultry in the past 12 months, and if so, how much input did you have on income generated from it?$^{M}$ | Yes, I had input into very few decisions | 0.0003 |
| Does anyone in your household currently have this chickens, ducks, turkeys or pigeons, and if so, who contributes most regarding a decision for a new purchase of it?$^{W}$ | Yes, spouse | 0.0003 |
| "[PERSON'S NAME] can't raise any livestock other than what she has. These are all that's available." Are you like this person?$^{W}$ | Yes/no | 0.0003 |
| Regarding the amount of sleep you got last night, was that:$^{W}$ | Less than average | 0.0003 |
| Has anyone in your household taken any loans or borrowed cash/in-kind from an informal lender in the past 12 months, and if so, who makes the decision on what to do with the money/item?$^{W}$ | Yes, self | 0.0003 |
| When decisions are made regarding what types of crops to grow, who is it that normally takes the decision, and if not you, to what extent do you think you can make decisions if you wanted to?$^{W}$ | Spouse, to a medium extent | 0.0003 |
| Has anyone in your household taken any loans or borrowed cash/in-kind from an informal lender in the past 12 months, and if so, who made the decision to borrow from them?$^{M}$ | Yes, spouse | 0.0003 |
| Your actions with respect to how to express your religious faith are motivated by a desire to avoid punishment or gain reward? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Somewhat true | 0.0003 |
| Do you yourself control the money needed to buy food from the market?$^{W}$ | Yes/No | 0.0003 |

Note: Random forest. Most important predictor variables ranked by feature importance. We report the survey question and categorical response that forms the predictor. Questions are edited for conciseness and clarity. Superscripts in the question column indicate whether the respondent is a man (M) or woman (W).

**Table A4:** Random Forest Time Use Variables

| Question | Answer | Feature Importance |
|---|---|---|
| Does anyone in your household currently have large livestock, and if so, who contributes most regarding a decision for a new purchase of it?$^{M}$ | Yes, spouse | 0.0022 |
| Your actions with respect to whether and how to express your religious faith are motivated by and reflect your own values and/or interests? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Somewhat true | 0.0021 |
| Your actions with respect to whether and how to express your religious faith are motivated by and reflect your own values and/or interests? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0016 |
| Your actions with respect to how to protect your self from violence are motivated by a desire to avoid punishment or gain reward? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{M}$ | Somewhat true | 0.0014 |
| Did you participate in wage or salary employment in the last 12 months, and if so, how much input did you have in decisions?$^{W}$ | Yes, input into most decisions | 0.0014 |
| Your actions with respect to what to do if you have a serious health problem are motivated by a desire to avoid punishment or gain reward? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{M}$ | Somewhat true | 0.0011 |
| Your actions with respect to how to protect yourself from violence are motivated by a desire to avoid blame or so that other people speak well of you? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{M}$ | Somewhat true | 0.0011 |
| Your actions with respect to how to protect yourself from violence are motivated by and reflect your own values and/or interests? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0010 |
| When decisions are made regarding major household expenditures of household life, who is it that normally takes the decision, and if not you, to what extent do you think you can make decisions if you wanted to?$^{M}$ | NA | 0.0010 |
| Did you participate in wage or salary employment in the last 12 months, and if so, how much input did you have on income generated from it?$^{W}$ | Yes, input into most decisions | 0.0010 |

Note: Random forest. Most important predictor variables ranked by feature importance. We report the survey question and categorical response that forms the predictor. Questions are edited for conciseness and clarity. Superscripts in the question column indicate whether the respondent is a man (M) or woman (W).

**Table A5:** Random Forest Health Variables

| Question | Answer | Feature Importance |
|---|---|---|
| Your actions with respect to minor household expenditures are motivated by a desire to avoid punishment or gain reward? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0022 |
| Your actions with respect to what to do if you have a serious illness are motivated by a desire to avoid blame or so that other people speak well of you? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0019 |
| Your actions with respect to how to protect yourself from violence are motivated by a desire to avoid blame or so that other people speak well of you? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0017 |
| Does anyone in your household have a house? If so, who decides whether to sell it?$^{M}$ | No, NA | 0.0015 |
| Who decides how to spend money on housing?$^{W}$ | Yourself | 0.0013 |
| When decisions are made regarding minor household expenditures, who is it that normally takes the decision, and if not you, to what extent do you think you can make decisions if you wanted to?$^{M}$ | Self, NA | 0.0011 |
| Does anyone in your household have large consumer durables? If so, who decides whether to give it away?$^{M}$ | No, NA | 0.0009 |
| When decisions are made regarding major household expenditures, who is it that normally takes the decision, and if not you, to what extent do you think you can make decisions if you wanted to?$^{W}$ | Husband, not at all | 0.0007 |
| Who decides how to spend money on healthcare?$^{W}$ | Yourself | 0.0007 |
| Your actions with respect to whether or not to use family planning are motivated by a desire to avoid blame or so that other people speak well of you? Can you tell me whether it is entirely true, somewhat true, not very true or never true?$^{W}$ | Always true | 0.0007 |

Note: Random forest. Most important predictor variables ranked by feature importance. We report the survey question and categorical response that forms the predictor. Questions are edited for conciseness and clarity. Superscripts in the question column indicate whether the respondent is a man (M) or woman (W).