Using machine learning and qualitative interviews to
design a five-question women’s agency index*

Seema Jayachandran  Monica Biradavolu  Jan Cooper
Northwestern University  QualAnalytics  Harvard University

February 12, 2021

Abstract

We develop a new short survey module for measuring women’s agency by combining mixed-methods data collection and machine learning. We select the best five survey questions for the module based on how strongly correlated they are with a “gold standard” measure of women’s agency. For a sample of 209 women in Haryana, India, we measure agency, first, through a semi-structured in-depth interview and, second, through a large set of close-ended questions. We use qualitative coding methods to score each woman’s agency based on the interview, which we treat as her true agency. To identify the subset of close-ended questions most predictive of the “truth,” we apply statistical methods similar to standard machine learning except that we specify how many survey questions are selected. The resulting 5-question index is as strongly correlated with the coded qualitative interview as is an index that uses all of the candidate questions. We also considered a second “gold standard” measure of agency, a real-stakes choice between money for oneself or one’s husband. This lab game, however, did not measure agency cleanly in our setting. Thus, our preferred survey measure of agency is the one validated against qualitative interviews.

Keywords: women’s empowerment, psychometric testing, criterion validity

JEL Codes: J16, O12, D13, I32

*Corresponding author: seema@northwestern.edu. We thank Ambika Chopra, Anubha Agarwal, Sahiba Lal, Azfar Karim, Bijoyetri Samaddar, Vrinda Kapoor, Ashley Wong, Jacob Gosselin, and Akhila Kovvuri for excellent research assistance, and the Bill and Melinda Gates Foundation for funding the study. We also thank Markus Goldstein, Jessica Heckert, Hazel Malapit, Ruth Meinzen-Dick, Agnes Quisumbing, Anita Raj, Greg Seymour, and several seminar and conference participants for helpful feedback. The study received institutional review board approval from Northwestern University and the Institute for Financial Management and Research.
1 Introduction

Women’s agency — or their ability to make and act on their choices for their lives — is an important concept in research and policy related to gender equality. Many policies aim to increase women’s agency, as a means for them to improve their health, economic security, and so forth, or as an end in itself.

Agency is a psychological construct, or a postulated attribute, which unlike, say, height, is not directly observable (Cronbach and Meehl, 1955). It is also multi-faceted: It encompasses the many domains of one’s life including reproductive health, employment, and civic engagement, and has both an instrumental component and an intrinsic one. The complexity of agency makes it a challenge to measure quantitatively.

Nonetheless, in quantitative studies, researchers often want a measure of women’s agency for use in statistical analyses. When evaluating the effects of an intervention, a researcher might want to test for an increase in agency (i.e., use the measure as an outcome variable) or investigate whether women with more agency enjoy larger benefits from the intervention (i.e., use it for subgroup analysis). An accurate and precise measure of agency is important for these purposes. Otherwise, one does not know whether a lack of statistical evidence that a program improved women’s agency is due to the absence of an effect or to the inadequacy of the measure, for example.

While the complexity of agency suggests the need for a long survey module to measure it, in practice, researchers often are seeking a short module. In particular, if agency is a secondary focus of their study, they want a small set of close-ended survey questions that enables them to measure agency. A longer module obviously could elicit more information but would be costlier to implement, in terms of not just money but also respondents’ time.

The goal of this study is to develop a new short survey module that can be used to measure women’s agency. Our contribution lies in the innovative method we use to select the survey questions. We choose them based on criterion validity. That is, we evaluate our measure by testing its correlation with a “gold standard” measure of women’s agency (DeVon et al., 2007). We start with a “true” measure of agency and then select the best short set of questions to combine into an index based on how strongly correlated the index is with the “truth.” We carry out this exercise through mixed-methods data collection in

---

1It is not one of our study’s aims to contribute to how agency is conceptualized; we follow the literature in how we conceptualize it. We focus on a woman’s instrumental agency (“power to”) and intrinsic agency (“power within”), and specifically her agency within her household and marriage (Rowlands, 1997; Kabeer, 1999). The next section reviews the literature that conceptualizes agency.
Haryana, India, and then data analysis.

How well a psychometric (e.g., an agency index) captures the concept it is trying to measure is called its validity (Jose et al., 2017). One approach to assessing validity is content validity; this entails an assessment that a measure appears to accurately represent the construct, covering all of its facets. Qualitative methods are often used to assess the content validity of survey questions and ensure that the measures are locally meaningful and relevant (Camfield et al., 2009; Small et al., 2008; Kanbur and Shaffer, 2007; Rao, 2002; Shaffer, 2013).

Another type of validity is construct validity. The researcher uses theory to predict that a construct is related to another variable. For example, we might posit that women’s agency is related to another factor Z because Z increases women’s agency or women’s agency increases Z. Construct validity consists of assessing a measure based on its correlation with Z. An example of an agency index developed using construct validity is the Survey-based Women’s Empowerment Index (SWPER), which is a combination of Demographic and Health Survey questions chosen because of their strong correlation with gender gaps in health and education (Ewerling et al., 2017). The premise behind the measure is that women’s agency narrows these gender gaps, or when these gaps are narrower, women have more agency. An advantage of construct validity is we almost always have data on factors that might affect or be affected by a construct. The disadvantage is we are rarely certain that women’s agency causes or is caused by Z; it is theoretically possible that women’s agency and education gender gaps have a weak association, for example.

Our approach, criterion validity, also uses the correlation between the measure in question and another variable, but here the other variable is a second, “gold standard” measure of the same construct. There is, of course, no perfect or “true” measure of women’s agency. In practice, we create a validated women’s agency index by benchmarking it against a better

---

2Researchers might collect qualitative data from the study population as a first step and then use it to design new survey questions. Studies that take a qualitative-first approach have used various methods, including open-ended interviews (Camfield and Ruta, 2007; Woodcock et al., 2009), life histories (Quisumbing, 2011), participatory techniques (Hargreaves et al., 2007) and longer-term ethnographic engagement (Crede and Borrego, 2013; Jha et al., 2007; Ware et al., 2003). Qualitative methods are also used to improve the validity of proposed questions through open-ended debriefing techniques during piloting of the questions. Open-ended debriefing includes interviews and group discussions with respondents about how they understood the questions or asking them to think aloud as they answer them or having a panel of experts review the questions (Bowden et al., 2002; Durham et al., 2011; Latcheva, 2011; Cohen and Saisana, 2014; Greco et al., 2018).

3In addition to validity, another characteristic of a measurement tool is its reliability, or how consistent the value would be if the tool were used again in the exact same context with the same person. Test-retest correlation is one way to assess reliability, for example.
way of measuring women’s agency, one that provides richer and more accurate data. If a
ingericher measure exists, then why not always use it? Because doing so is often impractical.
The better measures we use are more time-intensive, skill-intensive, logistically complex, or
generous, and thus are infeasible to include in most large-N studies.

The primary “gold standard” measures we use are semi-structured interviews conducted
by trained qualitative researchers. We use qualitative coding methods to score each woman’s
agency as conveyed through the interview. These interviews provide in-depth and nuanced
data but require highly skilled staff to conduct and code them. We also collected a second
“gold standard” measure of agency based on a lab game. In the game, which we adopt
from Almás et al. (2018), each woman makes a real-stakes choice between money for herself
or her husband. This lab game adds logistical complexity and costs to the fieldwork, but
observed behavior (“revealed preferences”) might be less subject to social desirability bias
than survey responses. We used these two quite different “gold standards” out of recognition
that researchers likely differ in which they prefer, according to their methodological taste.

We conduct the lab game among 443 women and choose a subsample of 209 women for the
semi-structured interviews.

The third way we measure women’s agency is through close-ended survey questions.
We ask a long list of questions, drawing on existing survey instruments. Our objective at
the data collection stage was to be comprehensive and agnostic about which were the best
questions, and then to later use a data-driven approach to select the best ones. We use
statistical methods to choose the best five questions and then combine them into an index.
There is nothing special about five questions, but this length seems appropriate for survey
designers seeking a short module on agency. Another benefit of a short validated module
is that it could serve as a common set of questions used in all surveys. A researcher could
choose to include many other questions on agency too, but the common questions would
allow for better comparisons across data sets, without being too onerous to include.

The goal of our statistical analysis is to select the best close-ended questions to field
from among the many candidates. The statistical methods we use for variable selection
are similar to standard supervised machine learning techniques except that we constrain
the number of survey questions selected. The primary method we use is LASSO stability
selection, in which the selected questions are those chosen most frequently when LASSO
is repeatedly run on subsamples of the data (Meinshausen and Bühlmann, 2010). This
method has previously been used by Kshirsagar et al. (2017) to select a short set of survey
questions for a proxy-means test of household poverty, for example. A second method we use is backward sequential selection (Liu and Motoda, 2012). This method entails starting with the full set of survey questions and iteratively removing the one question that leads to the smallest decrease in the explanatory power of the index, stopping when the desired number of questions (in our case, five) remain.

Turning to our results, we find that the interviews seemed successful in measuring agency, and when we use them as the “gold standard,” both of the statistical methods described above arrive at an index of women’s agency that is strongly correlated with the interview score. Three out of the five questions selected by LASSO stability selection (among 64 candidates) are also chosen by backward selection. Both five-question indices are considerably more correlated with the “truth” than if we had chosen the subset of questions randomly. More strikingly, the five-question indices have more explanatory power than indices constructed from all 64 candidate questions, either their first principal component or a standardized index that averages them. Interestingly, the algorithm-selected questions are quite specific ones about decision-making in particular situations, rather than questions that ask the woman about her power in general.

The lab game was ineffective in measuring agency in our study. The premise of the measure is that a woman with less agency will more often choose money for herself because she would not have a say in how money given to her husband is spent. We do see this behavior, but we also see an opposing force: some women with very low agency never want money for themselves because they consider money to be men’s domain or are fearful of their husband finding out and becoming angry. The survey index obtained when we apply the statistical techniques is only weakly correlated with the lab game behavior, consistent with this putatively “true” measure actually being very noisy. We conclude that only the semi-structured interviews can be considered a “gold standard” in our setting. Another advantage of the qualitative interviews is that they cover many domains of agency, not just financial agency.

The primary contribution of our study is methodological: We introduce a novel mixed-methods way to develop a survey measure. Using mixed methods in the design of measurement scales is not new (Onwuegbuzie et al., 2010; Zhou, 2019). For example, Creswell and Clark (2017) describe a process of using qualitative methods to define a construct and then using quantitative methods to assess the scale once it is developed. What is new is our use of statistical methods to choose quantitative questions by treating the qualitative measure
of the construct as the “gold standard.” We refer to this new approach to survey module
design as MASI, for MAchine learning and Semi-structured Interviews.\textsuperscript{4}

We believe that selecting survey questions based on their statistical correspondence to
qualitative interview data is innovative and has applications beyond women’s agency. Many
complex concepts are best measured with open-ended questions, yet there is practical need
for close-ended measures of them. One could similarly apply MASI to create survey modules
for other such constructs, such as financial insecurity or cultural assimilation.

The second contribution of our study is the new short survey module and index for
women’s agency that we develop. Our study adds to the literature proposing measures of
women’s agency or empowerment, which we review in the next section. We believe that the
five-question module validated against semi-structured interviews is a valuable new resource
for measuring women’s agency in north India, and perhaps elsewhere. A natural direction
for future research is to replicate the study elsewhere to create short modules appropriate for
other contexts and to assess the extent to which the same questions are or are not selected
elsewhere. One could also apply our methods to select a “universal” module based on how
robustly it predicts qualitative interview scores across multiple contexts.

2 Related literature on women’s agency

2.1 The concept of agency

Agency is one aspect of women’s empowerment. Empowerment as defined by Kabeer
(1999) encompasses resources, agency, and achievement and refers to the process of acquiring
the ability to make choices. Contemporary notions of empowerment often build on Amartya
Sen’s capabilities approach, as elaborated by Nussbaum (1999), who highlights that dignity
and the freedom to actively determine one’s life are central to human beings.

As one aspect of empowerment, agency refers to the ability to make decisions and act on
one’s goals. Agency is often defined in a way that captures both an intrinsic characteristic
and something with external, instrumental value. To do this, many definitions of agency
reflect both an internal feeling of agency (sometimes defined as the ability to set goals,
where the setting of goals is a reflection of the intrinsic sense of agency); and the external
actions of pursuing goals, which is the instrumental aspect of agency (Donald et al., 2020).

Scholars have also highlighted that the conceptualization of women’s agency depends

\textsuperscript{4}Masi means maternal aunt in Hindi.
on the context, for example differing in more coercive settings. Individual actions must be viewed within social, economic, and cultural contexts, and there are multiplicities and hidden forms of women’s agency (Campbell and Mannell, 2016).

2.2 Measurement of women’s agency

There is an array of research on how to measure women’s empowerment and agency. Donald et al. (2020) and Laszlo et al. (2020) provide excellent overviews of this literature.

Recent proposed measurement tools include the Women’s Empowerment in Agriculture Index (WEAI) (Alkire et al., 2013) and PRO-WEAI (Malapit et al., 2019). WEAI is a set of survey questions that measures empowerment, agency, and inclusion in the agricultural sector (Alkire et al., 2013). It aggregates an individual’s empowerment across five domains and also measures women’s status relative to men in the household. The index was designed based on analysis of household survey data collected in Guatemala, Uganda, and Bangladesh, and the index has been applied in several other contexts subsequently. PRO-WEAI adapts the WEAI to measure empowerment brought about by agriculture projects (Malapit et al., 2019). The PRO-WEAI uses the same methodology as the WEAI and includes further indicators that are most likely to change over the course of a project’s duration. This adaptation of the WEAI employed qualitative data from key informants and project participants to inform the construction of the project-level WEAI measure.

Another proposed measure is SWPER, which was developed by analyzing responses to Demographic and Health Survey questions among partnered women in 34 African countries (Ewerling et al., 2017). A principal component analysis extracted core components that were analyzed for convergent validity and external validity. SWPER includes 15 questions that represent three dimensions of empowerment: attitudes to violence, social independences, and decision making. Another recent contribution is Maiorano et al. (2020), who introduce a choices-values-norms framework for measuring agency. Specifically on India, Kishor and Gupta (2004) adapt WEAI for nutrition, while Richardson et al. (2019) develop an index of National Family Health Survey questions using confirmatory factor analysis.

A different strand of the literature assesses current practices for measuring women’s agency. Donald et al. (2020) and Laszlo et al. (2020) highlight conceptual challenges and provide frameworks to guide measurement. Peterman et al. (2021) investigate how robust results are to different ways of constructing agency indicators from commonly-used survey questions. They conclude that current practices are often insufficient to capture women’s
decision-making and call for further measurement innovation.

3 Description of study site and sample

3.1 Selection of study site and sample villages

We selected Kurukshetra district in the Indian state of Haryana as the study site based on several considerations. We chose north India because of our knowledge of the context and because women’s agency is an important topic of study there. To match our team’s language skills, we restricted attention to Hindi-speaking areas. Within this narrowed set of possible sites, we chose Kurukshetra for practical reasons. First, we could draw on a pool of female surveyors who had worked on earlier studies conducted by J-PAL South Asia, the research organization through which our fieldwork was conducted. Second, the main town was large enough that we could recruit two lead research assistants from New Delhi who would be willing to be based there for several months. Third, Kurukshetra was within a few hours of New Delhi by car or train, which facilitated site visits by the principal investigators.

We worked backwards from our target sample size of 210 semi-structured interviews to determine how many villages within Kurukshetra to include in our sample. We were able to recruit two qualitative interviewers, and about 100 interviews each was the most they could conduct within the three months we had planned for the data collection. We wanted to complete data collection in a village within two or three days so that there would not be discussion among women about our study that might prime their answers. We expected each interviewer to conduct two to three interviews per day, which implied that our team should conduct about 10 interviews per village. We, thus, included 21 villages in our sample in order to complete 210 interviews.

We had a separate, larger team of surveyors that conducted the quantitative surveys and lab game. The quantitative team spent about the same number of days in each village, collecting data from twice as many women. The final sample size for that team was 443 women, of whom the 209 women who participated in a semi-structured interview are a subset.

We chose a random sample of 21 villages that were representative of Kurukshetra, with the selection stratified by village population, distance from the district headquarters, and the ratio of male to female literacy.5 We created a randomly ordered list of potential sample

5Using the 2011 Census, of the 407 villages in Kurukshetra district, we excluded the top and bottom 5% of villages based on population, distance to the district headquarters, child sex ratio, and female literacy
villages that included more than 21 villages. We then visited the villages to obtain a list of households with young children from the village ASHA, or Accredited Social Health Worker. We planned to use these lists to choose households for the sample. In the few cases where we could not obtain a list from the ASHA, we replaced the village with the next village from its stratum on our list. Figure 1 shows the location of Kurukshetra district within India and the location of the 21 study villages.

3.2 Selection of study participants and descriptive statistics

We used the ASHA lists to choose a preliminary sample of households in each village. Our eligibility criterion was that a participant was a married woman with a child under the age of 10. We wanted the sample to be homogeneous in this way so that we could ask everyone similar questions, for example about their relationships with their husbands and about decisions over children’s health. The ASHA data included a household roster but not relationships among household members, so we chose households with a child under age 10 and a woman at least 15 years older than that child, who was feasibly the child’s mother. We aimed to enroll 20 women (and at most one per household) in the study, and we randomly chose 50% of them for the semi-structured interview.

We collected the data between February and May 2019. We varied whether the qualitative or quantitative data collection came first. The quantitative team started fieldwork in a random half of villages, and the qualitative team started in the other half; halfway rate. We also excluded a few villages with similar names as each other to avoid confusion in the field. Among the remaining 303 villages, we picked 2 or 3 villages in each of eight strata, defined by being above or below median population, distance to district headquarters, and ratio of male to female literacy.
through the data collection, they switched villages. (We do not find significant differences in measured agency, either qualitative or quantitative, based on the order of data collection.6

The first step when the first team visited a household was to verify the woman’s eligibility for the study, which also required that she speak Hindi. We then explained the study and obtained informed consent from women who chose to enroll in the study.

Table 1 reports summary statistics for the sample, based on data collected in the quantitative survey. The women are on average 30 years old with a youngest child who is five years old. Age of marriage averages 20.4 years, and women are 3 years younger than their husbands. The average years of schooling is 10. Most of the sample is Hindu, with Sikhism being the second most common religion. About a third of the sample belongs to a scheduled caste or scheduled tribe, and about half belong to an “other backward caste.” Only about a quarter of women are employed, consistent with the India-wide female employment rate.

4 Measuring agency with three types of data

4.1 Quantitative surveys

We administered a survey that asked close-ended questions to the full sample of 443 study participants. This survey, which was on average 45 minutes long, was conducted by female enumerators who had the typical qualifications for J-PAL South Asia quantitative surveyors.

After asking a few questions on demographic characteristics such as age and religion, the questionnaire focused on measures of women’s agency within her household. We ask a long list of close-ended questions measuring agency, aiming to be exhaustive. We drew on existing questions to measure agency used in other surveys. These included questions from the Demographic and Health Surveys, Relative Autonomy Index (Ryan and Deci, 2000; Vaz et al., 2016), a J-PAL toolkit on measuring women’s agency that aggregated survey questions that were used in several research studies (Glennerster et al., 2018), and the Sexual Relationship Power Scale (Pulerwitz et al., 2000). We also included a handful of questions that we developed ourselves.

Concatenating all of the existing modules would introduce a lot of redundancy, resulting

---
6There were some cases where a woman declined to participate in the second part of the data collection or the second team could not locate her. The sample size of 209 qualitative interviews that we focus on are those for whom we also have quantitative data. We also conducted interviews with another 9 women for whom we were not able to then collect quantitative data.
Table 1: Descriptive statistics for the sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full sample</th>
<th>Sample with qual. interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of respondents</td>
<td>443</td>
<td>209</td>
</tr>
<tr>
<td>Age</td>
<td>29.720</td>
<td>29.512</td>
</tr>
<tr>
<td>[4.953]</td>
<td>[4.778]</td>
<td></td>
</tr>
<tr>
<td>Age at marriage</td>
<td>20.377</td>
<td>20.316</td>
</tr>
<tr>
<td>[2.584]</td>
<td>[2.708]</td>
<td></td>
</tr>
<tr>
<td>Husband-wife age gap</td>
<td>2.946</td>
<td>2.914</td>
</tr>
<tr>
<td>[2.821]</td>
<td>[2.702]</td>
<td></td>
</tr>
<tr>
<td>Age of youngest child</td>
<td>4.989</td>
<td>5.019</td>
</tr>
<tr>
<td>[2.765]</td>
<td>[2.792]</td>
<td></td>
</tr>
<tr>
<td>Can read and write</td>
<td>0.986</td>
<td>0.986</td>
</tr>
<tr>
<td>[0.116]</td>
<td>[0.119]</td>
<td></td>
</tr>
<tr>
<td>Years of education</td>
<td>9.916</td>
<td>10.024</td>
</tr>
<tr>
<td>[3.258]</td>
<td>[3.175]</td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>0.165</td>
<td>0.182</td>
</tr>
<tr>
<td>[0.371]</td>
<td>[0.387]</td>
<td></td>
</tr>
<tr>
<td>Hindu</td>
<td>0.840</td>
<td>0.837</td>
</tr>
<tr>
<td>[0.367]</td>
<td>[0.370]</td>
<td></td>
</tr>
<tr>
<td>Sikh</td>
<td>0.151</td>
<td>0.144</td>
</tr>
<tr>
<td>[0.359]</td>
<td>[0.351]</td>
<td></td>
</tr>
<tr>
<td>Scheduled Caste/Scheduled Tribe</td>
<td>0.341</td>
<td>0.335</td>
</tr>
<tr>
<td>[0.475]</td>
<td>[0.473]</td>
<td></td>
</tr>
<tr>
<td>Other backward castes</td>
<td>0.501</td>
<td>0.502</td>
</tr>
<tr>
<td>[0.501]</td>
<td>[0.501]</td>
<td></td>
</tr>
<tr>
<td>Pukka house</td>
<td>0.386</td>
<td>0.373</td>
</tr>
<tr>
<td>[0.487]</td>
<td>[0.485]</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Table reports variable means and standard deviations.
in an overly long and repetitive survey from the respondent’s point of view, so we made judgment calls in removing questions that overlapped. In total, we asked 64 questions measuring agency. (The list of questions is included at the end of the appendix.)

Some of the agency questions were about the woman’s say in specific decisions, such as, “If money is available, who in your household decides whether to pay school fees for a relative from your side of the family?” and “Can you go unescorted to the next village?” Other questions were more general, asking the woman about her overall impression of her agency. An example is, “This is a ten step ladder, where on the bottom, the first step, people who are completely coerced or powerless stand, and on the highest step, the tenth step, stand those with the most ability to advance goals that they value in their own homes and in the world. On which step are you today?”

We convert each of the survey responses to a single numerical variable. Some of the responses have a natural numerical unit (e.g., days) or are binary. For questions asked on a Likert scale, we treat the categorical response as a cardinal variable. In a handful of cases where the numerical mapping is less clear, we make judgment calls. For example, in questions asked about whether women make decisions alone, jointly with their husband, or not at all, we code those responses as 2, 1, and 0. Note that we code all of the variables so that a higher value corresponds to more agency.

It is also possible to include multiple variables, or recodings, per survey question; the constraint for data collection is the number of survey questions. For the ladder question mentioned above, we could use variables for the response being ≥ 2, being ≥ 3, and so forth up to the response equaling 10. This approach uses more information, and we use it as a robustness check. We choose to use one variable per survey question in our main index for simplicity.⁷

### 4.2 Semi-structured interviews

The semi-structured interviews were on average 45 minutes long, ranging from 25 to 70 minutes. They were conducted primarily by two female interviewers who had prior experience with in-depth interviewing. A third interviewer conducted a few of the interviews. As training, one of the authors (MB) observed each interviewer conducting a few pilot interviews and provided feedback to improve interview skills. The interviewers and MB met weekly to discuss substantive and methodological issues that arose, with learnings fed back into

---

⁷A few of the questions have valid missing responses due to skip patterns in the survey. To include these questions in our analysis, we impute the value with the sample mean.
subsequent interviews.

The interviews, which were recorded, followed an interview guide that was refined through piloting. The initial interview guide covered five domains of agency within the household: the respondent’s decision-making around her children’s education and health, household expenditures, and her own fertility and mobility. In pilot interviews, employment emerged as another theme and was added to the interview guide as a sixth domain. The interviewers were trained to follow the interview guide, but to use their judgment to ask follow-up questions or diverge from the guide.

The open-endedness of interviews and the multiple domains allowed women to discuss direct and hidden strategies and the meanings behind the actions, including “bargaining and negotiation, deception and manipulation, subversion and resistance, and more intangible, cognitive processes of reflection and analysis” (Kabeer, 1999, p. 438).

To ensure privacy during the interviews, we paired each interviewer with someone initially recruited for our quantitative surveyor team who acted as a “distractor.” The distractor would have a discussion with other family members in a separate room so that the qualitative interviewer and study participant could have an uninterrupted private conversation.

The interviews were transcribed, and two people, the same two people who conducted the interviews, coded them using Dedoose software. We randomly assigned which interviews each person coded, so in about half the cases, it was an interview she had conducted.

We used a two-step approach to coding, following Deterding and Waters (2018). The first step in their “flexible coding” process is the development of “index codes” to represent the broad topics pursued during the research. In this study, the index codes were the six domains of agency included in the interview guide. The second step is the application of “analytic codes,” which emerge in the second reading of the transcripts. We paid attention to “speech practices” in our transcripts following Madhok (2014), since agency is often more than observable action and women’s own words open up the range of possibilities of what they consider agentic in their particular context.

The analytic codes were used to arrive at ranks (i.e., scores) and ranking definitions for each index code. MB and the coders triply coded and then discussed ten transcripts to harmonize how the coders interpreted and applied the codes.

The ranks ranged from 1 for a woman with the lowest level of agency to a 4 for a woman with the highest level of agency.\footnote{As we developed the coding approach by triply coding ten transcripts, we tried using a scale of 1 to 3, 1 to 4, and 1 to 5. We chose 1 to 4 because it seemed to best to capture the nuances in the interviews and}
the analytic codes were used to arrive at the definitions, in the mobility domain, a woman coded as a 1 needs explicit permission to leave the house and always goes accompanied by her husband or someone else to locations that are either inside or outside the village, which includes the neighborhood store, her children’s school, the hospital, the market, the bank and her natal village. If a woman has those restrictions but objects to them or sometimes tries to resist them, she is coded as a 2. That is, if the analytic codes “never goes alone” but also “resistance” were coded in the transcript under the index code “mobility,” the woman’s rank moved from 1 to 2. A woman who has some but not all of the restrictions was coded as a 3; for example, a woman may be allowed to go to locations inside the village by foot, but not be able to go unaccompanied to locations that require transportation. Women with the most agency over their mobility were coded as 4. They are able to go unaccompanied to all locations (the school, hospital, their natal village, etc.).

The one domain not initially coded on a 1 to 4 scale is fertility. Many women had allow us to define each rank distinctly.
Figure 3: Distribution of overall scores from semi-structured interviews

Notes: The histogram shows the overall coding of agency for women in the sample, which is the simple average across her scores in each of the six domains.

discordant levels of agency across the four sub-domains of number of children, birth spacing, reversible birth control, and sterilization, so we coded a woman separately in each of the sub-domains and then averaged these scores. This fertility score was then normalized to also range from 1 to 4. Figure 2 shows histograms of the domain-specific scores.

We then calculate an overall agency score for the woman as the average across the six domains. Figure 3 shows the distribution of the overall agency score, as coded from the semi-structured interviews. Hereafter, we refer to the overall score as the qualitative score.

4.3 Lab game

We used a “lab-in-the-field” game to measure women’s agency over household income. The game was conducted during the same visit and by the same surveyor as the quantitative survey. It took place in private at the end of the survey and took on average 15 minutes.

The measure uses real-stakes choices the woman makes, specifically her willingness to pay (WTP) to be the recipient of money given to the household. This measure was developed by Almås et al. (2018) in a study in urban Macedonia and has since been used in other settings (Barr et al., 2020). A potential advantage of a real-stakes choice is that it provides an objective, quantitative measure of the woman’s behavior. Because money is at stake, a respondent might be less subject to experimenter demand effects through which she gives insincere answers.

9We test robustness to creating a standardized index across the six domains in section 6.
The woman is offered a choice between ₹300 (4 USD) for herself or an amount of money $X$ to be given to her husband. We begin with $X = ₹700$. If the woman chooses the money for herself, we stop. If she chooses money for her husband, we decrement $X$ by ₹100 and ask again: Does she prefer ₹300 for herself or ₹600 for her husband? The last amount we ask about is $X = ₹100$. We inform her that any transfer of money to her will take place privately and that we will not communicate with her husband about the game if she chooses money for herself. If she chooses for her husband to get the money, we will give it to him and explain that it is tied to his wife’s participation in our study. We also let her know that we will implement one of her choices, selected at random. This procedure gives her an incentive to report her true preferences (Becker et al., 1964).

As Almås et al. (2018) write, in a unitary household, that is, if the husband and wife have identical preferences or are perfectly altruistic toward each other, “women should not be willing to pay anything in order to receive the transfer themselves, and should instead try to maximize the transfer amount.” But, the authors continue, “in a non-unitary model, the weaker the position of the woman in the household (the lower her control of resources), the more she should be willing to pay to obtain control of that transfer.” Some women might prefer ₹300 for themselves over ₹400 or even ₹700 for their husband because they would have so little say in how their husband’s money is spent. We calculate a woman’s WTP to be the recipient of the money, which is the highest amount $X$ at which she prefers money for her husband. If the highest amount is ₹700, her WTP is ₹400 (700 minus 300), for example. The higher the WTP, the more she is willing to forgo in total household money to be the recipient of the transfer, so the lower her agency. Thus, her WTP is an inverse measure of her agency. We use -$WTP$ as the variable measuring agency.

However, the measure did not work as theoretically intended. Many women always preferred that their husband get the money even when it was a smaller amount than ₹300. For example, they prefer ₹100 for their husbands to ₹300 for themselves. These women have a negative WTP to control financial resources. Figure 4 shows the distribution of WTP.

We debriefed with women who had a negative WTP to understand their behavior in the game (Jackson, 2011). This revealed that their choice was linked to having low agency; they believed that women should not get involved with household finances, or they feared that their husband would find out they received money and become angry. The theoretical premise of the measure is that low-agency women will have a higher demand for agency, but many women with low agency in fact did not want more agency.
Figure 4: Distribution of women’s WTP to be recipient of money in lab game

Notes: The figure is a histogram of women’s crossover point in the lab game, or the maximum amount they would forgo for their household to be the recipient of the money. A woman whose WTP is ₹400 prefers ₹300 for herself to ₹700 for her husband. A negative WTP means the woman prefers money to go to her husband, all else equal, e.g., -₹200 means that a woman prefers ₹300 for herself to her husband getting ₹100, but prefers ₹200 for her husband to ₹300 for herself.

As a partial fix for this problem, we bottom code WTP at 0. In other words, our measure of agency is min{−WTP, 0}. More importantly, after noticing this pattern in the field and then seeing the distribution of WTP, we became pessimistic that using WTP as the “truth” would yield a reliable survey measure of agency.

5 Statistical methods to choose best survey questions

The goal of our data analysis is to choose the five best close-ended survey questions to measure women’s agency. We do so by selecting those that are the best predictor of “true” agency.

An intuitive approach to selecting the survey questions would be to try every possible combination of five questions and choose the set that yields the highest $R^2$ when the true measure, $A$ is regressed on the variables, using ordinary least squares. This approach is subject to over-fitting, so we apply two other methods that have the same objective of selecting a best prediction but that are less subject to over-fitting the data.

Standard supervised machine learning techniques like LASSO and random forest share our goal of prediction without overfitting. The distinction here is we want to put a rigid constraint on the number of predictors to select. If LASSO chooses 15 variables, that would yield a survey module that is impractical for many purposes. Thus, the two statistical
algorithms we implement, described below, are similar to standard machine learning except
that they specify how many variables to select.

5.1 LASSO stability selection

The primary method we use is LASSO stability selection, in which the best questions
are those most commonly selected when LASSO is repeatedly run on subsamples of the data.
Meinshausen and Bühlmann (2010) show that variable selection through this combination of
regularized regressions (e.g., LASSO) and resampling (e.g., drawing subsamples) is less sen-
sitive to the choice of the tuning or regularization parameter than simply using a regularized
regression method such as LASSO.10

Kshirsagar et al. (2017) recently applied the stability selection technique with a similar
goal as ours of developing a short set of survey questions to measure something often mea-
sured with many questions. They select ten questions to measure a household’s likelihood of
being poor; their “true” measure of poverty is based on a longer survey module that collected
data on household consumption.

We use 50% subsamples and run LASSO 1000 times:11

1. Draw a 50% subsample of observations without replacement.

2. Run a LASSO regression of $A$ on all of the survey variables, keeping track of which
predictors are selected.12

3. Complete 1000 iterations of steps 1 and 2.

The proposed survey module consists of the 5 survey questions chosen most frequently
by LASSO across the iterations. We then combine them into an index by normalizing
each of the variables to have a standard deviation of 1 and mean of 0 and then averaging
the standardized variables. We refer to this type of aggregation as a standardized index.
Using the regression coefficients as weights to create a weighted index is another natural

10 As a brief primer on LASSO, it is a type of regularized regression. A regularized regression differs from
a standard regression in that the estimator “shrinks” some coefficients to zero to avoid the model over-fitting
the data. Thus, starting from a large set of regressors, only a subset have non-zero coefficient estimates,
or are selected for inclusion in the model. The tuning parameter specifies how much to penalize additional
variables in the model, that is, how aggressive the procedure should be in shrinking coefficients to zero.
11 Implemented in Stata on a standard desktop computer, the procedure takes 19 minutes to run. Backward
selection takes a few seconds.
12 The tuning parameter of the LASSO regression is chosen within each iteration by 10-fold cross-
validation.
way to combine the variables.\footnote{In the application by Kshirsagar et al. (2017) to a survey-based proxy means test of household poverty, after selecting variables, they again use a penalized regression to guide how to create the predicted poverty measure; they use the regression coefficients as weights.} We opt for just an average of the standardized variables for simplicity and to make the aggregation less dependent on the estimates.

Unlike in some prediction exercises, there is a “correct” sign of each regression coefficient in our case. The premise of our criterion validation exercise is that we are regressing one measure of agency on another, so the sign of the coefficients should be positive. Nothing in the statistical procedure guarantees or constrains the coefficients to be positive. Thus, one diagnostic for how well the procedure works is whether any of the coefficients are wrong-signed.

5.2 Backward sequential selection

The second method we use is an adapted version of the backward sequential selection technique (Liu and Motoda, 2012). We start with the full set of survey questions and iteratively remove the one that leads to the smallest decrease in the $R^2$ of an ordinary least squares regression, stopping when the target number (in our case, five) questions are left.\footnote{One can also run sequential selection in the forward direction, starting with an empty set and then sequentially adding the most predictive variable among the candidates. Backward selection typically outperforms forward selection (Leslie et al., 2018).} This method has been used previously by Leslie et al. (2018), for example, to select the most useful ten questions within the Service Provision Assessment, an international survey of health care facilities. In their application, they use the 649-item index as the benchmark and choose the best questions based on reliability, or how consistently their index classifies facilities into quintiles in a cross-validation exercise.

In implementing backward selection, all of the candidate variables could be included in a multivariate regression. Because ultimately most researchers will want to use them in one index, we combine the variables into an index at the selection stage. In other words, if we are at the iteration with $k$ variables left, instead of regressing $A$ on $k$ variables, we combine the $k$ variables into a standardized index and estimate a univariate regression.

The first step is to combine all the candidate survey variables on agency (64 in our case) in an index. Then we iteratively remove variables as follows.

1. Discard one of the available variables and combine the remaining $k$ variables into an index (after normalizing them).

2. Calculate the $R^2$ when $A$ (“true” measure of agency) is regressed on the index.
3. Repeat steps 1 and 2 for all remaining variables.

4. Drop from the list the variable that led to the smallest loss of $R^2$, relative to including all $k$ in the set.

5. Repeat steps 1 to 4 until the desired number of variables for the index is reached.

The last five questions that remain comprise the proposed survey module, and the standardized index based on them is the proposed measure of women’s agency.

6 Results: Validated survey module for women’s agency

6.1 Based on semi-structured interviews as gold standard

We report the best set of survey questions to measure agency, as determined by our criterion validation method, in Table 2. These are the questions chosen based on their correspondence with the qualitative score. Overall, both statistical methods select a similar set of five questions that capture a large amount of the variation in the qualitative score.

Table 2, column (1) reports the questions selected with our preferred statistical technique of LASSO stability selection. The numbers in the cells are the rank for the question, in terms of how often it was selected. The top question is about decision-making about large household purchases like a cow or bicycle. The variable was selected in 87% of the LASSO iterations, as reported in Table 3. The fifth question was selected 59% of the time. Table 3 provides the frequency of selection for the top ten variables; if a researcher is seeking a ten-question module, these are the best choices based on the algorithm. We find that the fifth- to seventh-ranked questions perform quite similarly, and the biggest gains from the algorithmic approach seem to be from identifying the best three questions. The least often one of the 64 candidate questions was selected by LASSO was 3% of the time.

Interestingly, none of the general questions that ask a woman to assess her overall agency or perception of her power are selected. The top three questions ask about the woman’s role in specific purchase decisions: large household purchases, clothing for the woman, and items in the market. The other two questions pertain to agency over her physical mobility, specifically whether she can visit women in her neighborhood without permission, and her children’s health care. The mobility question highlights that the best five-question module is likely to differ by context; restrictions on women’s travel within her village are more common in India than many other places (Jayachandran, 2015; Naybor et al., 2016).
Table 2: Selected survey questions using semi-structured interviews

<table>
<thead>
<tr>
<th>Question</th>
<th>(1) LASSO stability selection</th>
<th>(2) Backward selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion heard when expensive item like a bicycle or cow is purchased?</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Need permission from other household members to buy clothing for self?</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Allowed to buy things in the market without asking partner?</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Are you permitted to visit women in other neighborhoods to talk with them?</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Who do you consult with for decisions regarding your children’s health care?</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Allowed to go alone to meet your friends for any reason?</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Who in hhld decides to pay school fees for a relative from your side of family?</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

5-Question Index $R^2$                  0.2888 0.2867
5-Question Multivariate Regression $R^2$ 0.2896 0.2893

Notes: The table lists the top 5 survey questions selected. The numbers in the cells in columns (1) and (2) indicate the selection order, with 1 referring to the best, or most predictive question.
Table 3: Frequency of variable selection with LASSO stability selection (using semi-structured interviews)

<table>
<thead>
<tr>
<th>Question</th>
<th>Percent of times selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion heard when expensive item like a bicycle or cow is purchased?</td>
<td>86.8</td>
</tr>
<tr>
<td>Need permission from other household members to buy clothing for self?</td>
<td>76.8</td>
</tr>
<tr>
<td>Allowed to buy things in the market without asking partner?</td>
<td>75.7</td>
</tr>
<tr>
<td>Are you permitted to visit women in other neighborhoods to talk with them?</td>
<td>62.8</td>
</tr>
<tr>
<td>Who do you consult with for decisions regarding your children’s health care?</td>
<td>59.4</td>
</tr>
<tr>
<td>Allowed to go alone to meet your friends for any reason?</td>
<td>58</td>
</tr>
<tr>
<td>Are you permitted to visit any place riding on public transport?</td>
<td>57.4</td>
</tr>
<tr>
<td>Can decide by self to purchase emergency medicine for child</td>
<td>55.4</td>
</tr>
<tr>
<td>When husband has different opinion, voice opinion and argue more often than voice opinion but do as he says</td>
<td>53.2</td>
</tr>
<tr>
<td>In last 12 months, how often you and husband discussed children’s expenses</td>
<td>50.8</td>
</tr>
</tbody>
</table>

Notes: The numbers reported are how often, out of 1000 iterations of LASSO on 50% subsamples, a variable was selected as a regressor in the LASSO stability selection procedure. The dependent variable is the semi-structured interview score.
All of the selected variables are predictive in the correct direction; that is, with the variables coded such that, theoretically, a higher value represents more agency, the raw correlation with the qualitative score is always positive. Appendix Table 1 shows the correlation between the qualitative score and each of the selected variables.

The proposed way to combine the survey questions into one measure is as the average of five variables: We encode each survey question as a continuous variable, make the variables comparable by normalizing each to have a standard deviation of 1, and then average them. At the bottom of the table, we report the $R^2$ of a univariate regression of the qualitative score on the survey-based index. The $R^2$ of 0.29 corresponds to a correlation coefficient of 0.54. The next row shows what the $R^2$ is if we instead regress the qualitative score on the five survey variables separately. The similar $R^2$ indicates that one does not lose much information by using the simple average. The simple average, thus, seems preferable to more complicated and less transparent ways of aggregating the variables, such as taking a weighted average using the multivariate regression coefficients as the weights.

Appendix Table 2 shows the correlation between the survey index and qualitative scores in each of the six domains. The index is most strongly correlated with the household expenditures and mobility domain, which is unsurprising, as four of the five selected questions are within those two domains.

Table 2, column (2) reports the top five questions selected using backward selection. Three of them are in the set chosen by LASSO stability selection, though not in the same order. The new variables selected pertain to household spending and mobility, specifically paying for education fees for the woman’s relatives and visiting friends alone. The $R^2$ using this index is 0.29, very similar to the LASSO stability selection $R^2$.

We calculated the qualitative score by averaging the six domain-specific scores. To understand how sensitive the selection of questions is to this choice, we repeated the procedure using an alternative qualitative score that is a standardized index across the domains; that is, we normalized each domain score to have a standard deviation of 1 before averaging them. This change did not alter the top five questions selected using either LASSO stability or backward selection. While this amount of insensitivity need not always hold (for example, the tenth best question according to LASSO stability selection is sensitive to this choice), it does provide some additional reassurance about the robustness of our method.
Comparison to randomly choosing variables

One way to gauge how valuable it is to use an algorithmic approach to survey question selection is to compare it to ad hoc selection. Figure 5 plots a histogram of index performance – specifically the $R^2$ when the qualitative score is regressed on the index – if we randomly select five questions from among the 64 candidates. The median $R^2$ with randomly sets of variables is 0.06. The two algorithm-selected indices have explanatory power that is well above not just the median, but even the 99.9th percentile of the distribution using randomly selected variables.

Figure 5: Selected indices compared to five randomly chosen variables

Notes: We take 1000 random draws of 5 out of 64 questions. The figure plots the distribution of $R^2$ when the qualitative score is regressed on a standardized index combining the 5 variables.

Comparison to LASSO

When we estimate standard LASSO using the qualitative score as the dependent variable and the 64 candidate survey variables as potential regressors, LASSO selects 19 regressors. (The selected regressors are listed in Appendix Table 3.) Reassuringly, all 7 survey questions chosen across our two methods are among them. This need not have been the case: backward selection does not use LASSO, and LASSO stability selection is estimated on subsamples. (In contrast, when we use the lab game behavior as the dependent variable, LASSO does not choose the selected survey questions.)

If all of the LASSO-selected variables are combined into a standardized index, the $R^2$
is 0.27. Thus, one does not lose explanatory power using our 5 questions versus a longer module of 19 questions. Of course, one can also use the predicted value of the regression as the agency index. Doing so yields an $R^2$ is 0.44, so considerably higher than one achieves with the five-question index, but at the cost of a longer survey module.

**Comparison to using all 64 close-ended survey questions**

Another benchmark is if we used all 64 quantitative variables. The $R^2$ of a multivariate regression using all of them is 0.53. In other words, all the variables collectively explain 53% of the variation. With five questions combined in an index, one explains 27% to 29% of the variation. Thus, we sacrifice less than half of the explanatory power when we use just 5 out of 64, or 8%, of the potential survey questions, and combine them into one measure.

Using all 64 variables in a standardized index actually leads to a lower $R^2$ of 0.20 than with the index based on five variables. The cost of using more variables is not just that it requires a longer survey, but also that some variables are weak predictors of agency (as measured by the interview), so including them lowers the overall explanatory power of the index.

An alternative way to extract information from the 64 variables is to use principal component analysis. If we use the first principal component of all the variables as the measure of agency, in a regression of the qualitative score on the principal component, the $R^2$ is 0.23, which is again lower than what our algorithms achieve.

**Tradeoff between length of module and explanatory power of index**

The fact that an index using all 64 survey variables performs worse than using the five selected variables raises the question of how index performance is related to the number of variables selected. We repeated the LASSO stability and backwards selection algorithms varying the number of selected variables from 1 to 64. Figure 6 plots the explanatory power of the selected indices. For backward selection, the $R^2$ peaks at 0.35, with the best 12 questions included. Recall that the best 5 questions yield an $R^2$ of 0.28. Using backward selection, the maximum $R^2$ is achieved with 16 questions and is at 0.38.

Thus, there is a tradeoff between a shorter survey module and an agency index that captures more information, up to a point. A researcher willing to use a longer module could take the best 10 or 15 questions instead of the best 5 that we have focused on. But what is also apparent is that after a point, even if fielding a longer survey were not costly, using a larger number of agency variables in the index seems to hurt performance.
Figure 6: Explanatory power of selected indices when number of questions is varied

Panel (a): LASSO stability selection

Panel (b): Backward selection

Notes: The figures plot the $R^2$ of a regression of the qualitative score on an index constructed from the best $k$ variables selected by the algorithm; the value $k$ is plotted on the horizontal axis. The horizontal axis stops at 30 variables for LASSO stability selection because only 30 of the variables are ever selected in the LASSO iterations.
Correlation with characteristics often associated with women’s agency

Another way to potentially assess the selected indices is to see how correlated they are with factors often associated with agency. Three factors we collected data on that, a priori, one expects to be positively associated with agency are education, age at marriage, and the gap between the wife’s and husband’s age. Somewhat surprisingly, none of these factors are strongly correlated with the qualitative score itself, as reported in Appendix Table 2. The household expenditures domain seems to be driving this; less educated women, women who married earlier, and women with a larger husband-wife age gap have more financial agency within the household, as assessed through the semi-structured interviews. These patterns mean that these characteristics are less useful for testing the construct validity of the selected indices.

6.2 Based on lab game as gold standard

Given the problems with the lab game discussed in section 4.3, it is unsurprising that the statistical methods do not perform as well when the lab game is the “truth.” For completeness, we report the selected questions in Appendix Table 4. One indication that the selected questions are less reliable when we use the WTP measure is that the index combining survey questions is not strongly correlated with the “true measure” ($R^2$ of 0.05, compared to an $R^2$ of 0.29 with the qualitative interviews). A second stark result is that standard LASSO selects none of the top variables according to backward selection and LASSO stability selection, and in fact, no variables at all. None of the close-ended questions robustly predict behavior in the lab game. For these reasons, combined with the raw data and our observations from the field, we conclude that the lab game was an inadequate tool for measuring women’s agency in our context, or at least as implemented in our study.

7 Conclusion

One contribution of this study is to create a new survey module and index women’s agency, for use in north India and perhaps elsewhere. We select a 5-question survey module, from a starting set of 64 questions, in a data-driven way. Interestingly, the selected questions ask women about their agency in quite specific situations; more general questions are not selected. We propose as a women’s agency index the standardized average of the five variables mapping to the selected questions.

We believe this short module could be useful for two purposes, first, for those seeking
an off-the-shelf short way to measure agency and, second, as a set of standard questions that could be asked as part of longer modules, to enable comparisons across studies. The module was created using data from one district in India, so a valuable direction for future research is to replicate the study elsewhere.

The second contribution of our study is to introduce a new method for developing validated measures of constructs. The method combines machine learning and semi-structured interviews, or MASI for short. Based on the principle of criterion validation, we vet quantitative measures of a construct — women’s agency, in our case — by benchmarking them against semi-structured interviews, conducted and coded by trained qualitative researchers. Specifically, we use statistical techniques that build on or are akin to LASSO to select the best survey questions based on how well they correspond to the measure of agency obtained through the time- and skill-intensive in-depth interviews.

Behavior in a lab game we conducted mapped to agency in a nuanced way, making it unsuitable as a “gold standard” measure. The game uses a high demand for agency as a proxy for having low agency, but many women in our sample with very low agency did not want more agency. We therefore conclude that using qualitative data as the “true” measure of agency may be preferable. Semi-structured interviews have two other nice features. First, they can cover any number of domains, even beyond those we covered, e.g., political agency. Second, using qualitative interviews, one can easily apply our approach to create quantitative measures of concepts very different from women’s agency. We believe that combining machine learning and semi-structured interviews to develop short survey measures of complex constructs has many promising applications.
References


Camfield, L. and D. Ruta (2007). Translation is not enough: using the Global Person Generated Index (GPGI) to assess individual quality of life in Bangladesh, Thailand, and Ethiopia. *Quality of Life Research 16*(6), 1039–1051.


Jose, R., N. Bhan, and A. Raj (2017). EMERGE measurement guidelines report 2: How to create scientifically valid social and behavioral measures on gender equality and empowerment. Center on Gender Equity and Health (GEH), University of California, San Diego School of Medicine. San Diego, CA.


Malapit, H., A. Quisumbing, R. Meinzen-Dick, G. Seymour, E. M. Martinez, J. Heckert,


### App. Table 1: Correlation of qualitative score and selected survey variables

<table>
<thead>
<tr>
<th>Qual-interview agency score</th>
<th>Opinion heard when expensive item like a bicycle or cow is purchased?</th>
<th>Need permission from other household members to buy clothing for self?</th>
<th>Allowed to buy things in the market without asking partner?</th>
<th>Are you permitted to visit women in other neighborhoods to talk with them?</th>
<th>Are you permitted to visit any place riding on public transport?</th>
<th>Allowed to go alone to meet your friends for any reason?</th>
<th>Who in hhld decides to pay school fees for a relative from your side of family?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qual-interview agency score</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who do you consult with for decisions regarding your children’s health care?</td>
<td>0.218</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opinion heard when expensive item like a bicycle or cow is purchased?</td>
<td>0.318</td>
<td>-0.054</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allowed to buy things in the market without asking partner?</td>
<td>0.346</td>
<td>-0.006</td>
<td>0.287</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who in hhld decides to pay school fees for a relative from your side of family?</td>
<td>0.176</td>
<td>-0.013</td>
<td>0.019</td>
<td>0.206</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Need permission from other household members to buy clothing for self?</td>
<td>0.338</td>
<td>0.123</td>
<td>0.192</td>
<td>0.324</td>
<td>0.143</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Are you permitted to visit women in other neighborhoods to talk with them?</td>
<td>0.295</td>
<td>0.119</td>
<td>0.120</td>
<td>0.211</td>
<td>-0.139</td>
<td>0.155</td>
<td>1.000</td>
</tr>
<tr>
<td>Allowed to go alone to meet your friends for any reason?</td>
<td>0.280</td>
<td>0.190</td>
<td>0.091</td>
<td>0.218</td>
<td>-0.018</td>
<td>0.071</td>
<td>0.319</td>
</tr>
<tr>
<td>Qual-interview agency score</td>
<td>LASSO stability selection 5-Q Index</td>
<td>Backward selection 5-Q Index</td>
<td>Fertility score</td>
<td>Education score</td>
<td>Health score</td>
<td>HH Expenses score</td>
<td>Mobility score</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Qual-interview agency score</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LASSO stability selection 5-Q Index</td>
<td>0.535</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backward selection 5-Q Index</td>
<td>0.537</td>
<td>0.817</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fertility score</td>
<td>0.341</td>
<td>0.199</td>
<td>0.237</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education score</td>
<td>0.662</td>
<td>0.296</td>
<td>0.262</td>
<td>0.088</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health score</td>
<td>0.631</td>
<td>0.281</td>
<td>0.312</td>
<td>0.132</td>
<td>0.416</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HH Expenses score</td>
<td>0.710</td>
<td>0.441</td>
<td>0.437</td>
<td>0.097</td>
<td>0.364</td>
<td>0.363</td>
<td>1.000</td>
</tr>
<tr>
<td>Mobility score</td>
<td>0.696</td>
<td>0.338</td>
<td>0.358</td>
<td>0.005</td>
<td>0.306</td>
<td>0.357</td>
<td>0.526</td>
</tr>
<tr>
<td>Work score</td>
<td>0.467</td>
<td>0.313</td>
<td>0.283</td>
<td>0.140</td>
<td>0.156</td>
<td>0.027</td>
<td>0.079</td>
</tr>
<tr>
<td>Less than 8 years of education</td>
<td>-0.039</td>
<td>-0.094</td>
<td>-0.162</td>
<td>-0.166</td>
<td>-0.016</td>
<td>-0.028</td>
<td>0.145</td>
</tr>
<tr>
<td>Age at marriage</td>
<td>-0.064</td>
<td>-0.012</td>
<td>-0.020</td>
<td>-0.026</td>
<td>0.004</td>
<td>0.020</td>
<td>-0.177</td>
</tr>
<tr>
<td>Husband-wife age gap</td>
<td>0.053</td>
<td>0.019</td>
<td>0.039</td>
<td>0.017</td>
<td>-0.003</td>
<td>0.041</td>
<td>0.165</td>
</tr>
</tbody>
</table>
App. Table 3: All variables selected by regular LASSO (using semi-structured interview)

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can decide by self to purchase emergency medicine for child</td>
</tr>
<tr>
<td>Who accompanied you to healthcare provider?</td>
</tr>
<tr>
<td>Who do you consult with for decisions regarding your children’s health care?</td>
</tr>
<tr>
<td>Opinion heard when expensive item like a bicycle or cow is purchased?</td>
</tr>
<tr>
<td>Allowed to buy things in the market without asking partner?</td>
</tr>
<tr>
<td>Who in hhld decides to pay school fees for a relative from your side of family?</td>
</tr>
<tr>
<td>Who in household decides purchasing item like radio or paraffin lamp?</td>
</tr>
<tr>
<td>Need permission from other household members to buy clothing for self?</td>
</tr>
<tr>
<td>Do you have a bank or savings account that you yourself use?</td>
</tr>
<tr>
<td>Are you permitted to visit any place riding on public transport?</td>
</tr>
<tr>
<td>Are you permitted to visit women in other neighborhoods to talk with them?</td>
</tr>
<tr>
<td>Are you allowed to go alone to a relative’s house inside the village?</td>
</tr>
<tr>
<td>Allowed to go alone to meet your friends for any reason?</td>
</tr>
<tr>
<td>A wife should obey her husband, even if she disagrees.</td>
</tr>
<tr>
<td>In last 12 months, how often you and husband discussed domestic work</td>
</tr>
<tr>
<td>When husband has different opinion, voice opinion and argue more often than voice opinion but do as he says</td>
</tr>
<tr>
<td>In last 12 months, how often you and husband discussed children’s expenses</td>
</tr>
<tr>
<td>In last 12 months, how often you and husband discussed your relatives</td>
</tr>
<tr>
<td>On which step are you (0=powerless to 10=most ability)</td>
</tr>
</tbody>
</table>

Notes: The variables listed are the 19 ones chosen by standard LASSO when the dependent variable is the semi-structured interview score the possible regressors are the 64 close-ended survey questions.
## App. Table 4: Selected survey questions using lab game

<table>
<thead>
<tr>
<th>Question</th>
<th>(1) LASSO stability selection</th>
<th>(2) Backward selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>He (husband) expects me not to contradict him in public</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Since New Year’s Day, how often has this (husband threatening to hurt you) happened?</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>On which step are you (0=powerless to 10=most ability)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Are you allowed to go alone to a relative’s house inside the village?</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Opinion heard when expensive item like a bicycle or cow is purchased?</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Has partner/other male ever threatened to hurt or harm you</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

| 5-Question Index $R^2$       | 0.0455                        | 0.0465                    |
| 5-Question Multivariate Regression $R^2$ | 0.0474                        | 0.0478                    |

Notes: The table lists the top 5 survey questions selected. The numbers in the cells in columns (1) and (2) indicate the selection order, with 1 referring to the best, or most predictive question. * denotes that the variable, when coded so that a higher value maps to more agency, has a negative coefficient in a univariate regression with WTP as the dependent variable.
## List of all close-ended questions measuring agency used

<table>
<thead>
<tr>
<th>Question</th>
<th>Responses</th>
</tr>
</thead>
</table>
| Who do you consult with for decisions regarding your children’s health care?  
[a](#) | 1. No one, I decide on my own  
2. My husband  
3. Father-in-law  
4. Mother-in-law  
5. Relatives from my husband’s side  
6. Other (specify):  
7. Don’t know  
8. Refused to answer | |
| Imagine that you were home alone, without your (pre-fill xx response from previous question) and one of your children was very sick. Could you make the choice on your own to purchase medication to treat your child? | 1. Yes  
2. No  
3. Don’t know  
4. Refused to answer | |
| When was the last time you were unwell and visited a health care provider for treatment? | 1. [DD]  
2. [MM]  
3. [YY]  
4. Don’t know  
5. Refused to answer | |
| Who accompanied you to the provider?  
*Surveyor: Do not read out the response options aloud*  
[b](#) | 1. Went alone  
2. Husband  
3. Mother-in-law  
4. Sister-in-law  
5. Father-in-law  
6. Brother-in-law  
7. Son  
8. Father  
9. Brother  
10. Mother  
11. Sister  
12. Male relative  
13. Female relative  
14. Male non-relative  
15. Female non-relative  
6. Other (specify):  
7. Don’t know  
8. Refused to answer | |
| Getting permission to go? | 1. Big problem  
2. Small problem  
3. No problem  
4. Don’t know  
5. Refused to answer | |
| Finding someone to go with you? | 1. Big problem  
2. Small problem  
3. No problem  
4. Don’t know  
5. Refused to answer | |
| “When I am in a difficult situation, I can usually find my way out of it.” | 1. Strongly agree  
2. Somewhat agree  
3. Neither agree nor disagree  
4. Somewhat disagree  
5. Strongly disagree  
6. Don’t know  
7. Refused to answer | |
| “I feel helpless.” | 1. Strongly agree  
2. Somewhat agree  
3. Neither agree nor disagree  
4. Somewhat disagree  
5. Strongly disagree  
6. Don’t know  
7. Refused to answer | |
| “I feel I can provide for my family and meet my family’s needs.” | 1. Strongly agree  
2. Somewhat agree  
3. Neither agree nor disagree  
4. Somewhat disagree  
5. Strongly disagree  
6. Don’t know  
7. Refused to answer | |
| “I have no confidence in myself.” | 1. Strongly agree  
2. Somewhat agree  
3. Neither agree nor disagree  
4. Somewhat disagree  
5. Strongly disagree  
6. Don’t know  
7. Refused to answer | |
| When you have small amounts of money, such as 20 or 50 INR, can you decide how to spend it on your own? | 1. Yes, always  
2. Yes, usually  
3. Yes, sometimes  
4. Rarely  
5. Very rarely  
6. Never  
7. Don’t know  
8. Refused to answer | |
| When an expensive item like a bicycle or a cow is purchased by the household, is your opinion listened to in the decision of what to buy? | 1. Yes, always  
2. Yes, usually  
3. Yes, sometimes  
4. Rarely  
5. Very rarely  
6. Never  
7. Don’t know  
8. Refused to answer | |
| If you have some money you have earned, can you use it to purchase clothing for yourself or children without asking the permission of anyone else? | 1. Yes, always  
2. Yes, usually  
3. Yes, sometimes  
4. Rarely  
5. Very rarely  
6. Never  
7. Don’t know  
8. Refused to answer | |

---

[a](#) Recoded as: 1. Relatives from my husband’s side, Father-in-law, or mother-in-law, 2. My husband, 3. No one, I decide on my own

[b](#) Recoded as: 1. With male relative, 2. With female relative, 3. Went alone

(Continue on next page)
<table>
<thead>
<tr>
<th>Question</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you have to ask the permission of other household members to buy: Vegetables or fruits</td>
<td>1. Yes 2. No 3. Have never bought -99. Refused to answer</td>
</tr>
<tr>
<td>Do you have to ask the permission of other household members to buy: Clothing for yourself</td>
<td>1. Yes 2. No 3. Have never bought -99. Refused to answer</td>
</tr>
<tr>
<td>Do you have to ask the permission of other household members to buy: Medicines for yourself</td>
<td>1. Yes 2. No 3. Have never bought -99. Refused to answer</td>
</tr>
<tr>
<td>Do you have to ask the permission of other household members to buy: Personal supplies (soap, shampoo, dental paste, sanitary napkins, etc.)?</td>
<td>1. Yes 2. No 3. Have never bought -99. Refused to answer</td>
</tr>
<tr>
<td>How much money do you usually have on hand to meet these types of expenses?</td>
<td>1. Yes 2. No -98. Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>Do you have access to any cash available now for buying HH food or medicine if you suddenly needed something?</td>
<td>1. Yes 2. No -98. Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>Do you have a bank or savings account that you yourself use?</td>
<td>1. Yes 2. No -99. Refused to answer</td>
</tr>
<tr>
<td>Are you allowed to go alone to a relative’s house inside the village?</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
</tbody>
</table>
(Continuation of table listing quantitative survey questions)

<table>
<thead>
<tr>
<th>Question</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you allowed to go to the school/college alone or with friends?</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>Are you allowed to go alone to meet your friends for any reason (to get school notes, chat, play etc.)?</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>In the last one year, have you ever gone to the market within your village to buy personal items with friends? (no guardians)</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>In the last one year, have you ever gone to the market within your village to buy personal items alone?</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>In the last one year, have you ever attended any sort of community events/activities? (Ex: fair, theatre, cultural program, religious event)</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>In the last one year, have you ever attended one of these events without guardians present (either alone or with friends)?</td>
<td>1. Yes 2. No -97. Not applicable -99. Refused to answer</td>
</tr>
<tr>
<td>A woman should be able to choose her own friends, even if her husband disapproves</td>
<td>1. Strongly agree 2. Agree 3. Neither agree nor disagree 4. Disagree 5. Strongly disagree -98. Don’t know -99. Refused to answer</td>
</tr>
</tbody>
</table>

(Continue on next page)
(Continuation of table listing quantitative survey questions)

<table>
<thead>
<tr>
<th>Question</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Now I am going to read a list of things that might describe your current partner. Please tell me how closely this describes your current partner. G.12. Most of the time when we disagree about what to do, we do what my partner wants to do. G.13. My partner treats me well. G.14. My partner won’t let me wear certain things. G.15. When my partner and I are together, I’m pretty quiet. G.16. He expects me not to contradict him in public . G.17. He is upset if I express an opinion that disagrees with him G.18. [Follow up to previous] I often express my opinion when I disagree with my husband. G.19. My partner has more say than I do about important decisions that affect us. G.20. My partner tells me who I can spend time with. G.21. I feel trapped or stuck in our relationship. G.22. My partner does what he wants, even if I do not want him to. G.23. When my partner and I disagree, he gets his way most of the time. G.24. My partner always wants to know where I am.</td>
<td>1. Strongly agree 2. Agree 3. Neither agree nor disagree/Neutral 4. Disagree 5. Strongly disagree -98. Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>When your husband has a different opinion from you on a particular decision what do you do. Please tell us how many times you adopt each of these approaches (Surveyor: guide her through all the six options) G.26. - Don’t voice your opinion but do what you think is right G.27. - Don’t voice your opinion and wait for another occasion to see if he changes his mind G.28. - Don’t voice your opinion but do what your husband thinks is right G.29. Voice your opinion but also make it clear that you will go along with his view G.30. Voice your opinion and argue why your choice is better G.31. - Other: Please explain c</td>
<td>1. Never 2. Rarely 3. Sometimes 4. Often 5. Most of the times 6. All the time -97. Not Applicable -98. Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>When you disagree with your husband, does he get angry with you?</td>
<td>1. Yes 2. No -99. Refused to answer</td>
</tr>
<tr>
<td>Has he (your husband, or other adult male in your household) ever threatened to hurt or harm you or someone close to you?</td>
<td>1. Yes 2. No -99. Refused to answer</td>
</tr>
</tbody>
</table>

(cRecoded as four binary variables for yes/no questions: When husband has different opinion, voice opinion and argue more often than voice opinion but do as he says; When husband has different opinion, voice opinion but do as he says more often than not voicing opinion and waiting for him to change mind; When husband has different opinion, wait for him to change mind more often than do as he says (but don’t voice opinion); When husband has different opinion do what you think more often than what he says (but don’t voice opinion)
<table>
<thead>
<tr>
<th>Question</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s wrong for me to question people who are in charge or in authority, like teachers or parents, leaders in the village etc.</td>
<td>1. Strongly agree 2. Agree 3. Neither agree nor disagree 4. Disagree 5. Strongly disagree -98. Don’t know 99. Refused to answer</td>
</tr>
<tr>
<td>Do you feel that people like yourself can generally change things in your community if they want to?</td>
<td>1. Yes, very easily 2. Yes, fairly easily 3. Yes, but with a little difficulty 4. Yes, but with a great deal of difficulty 5. No, not at all -98, Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>Do you participate in making major household purchases? Major household purchases include things like refrigerator, television, vehicle, livestock etc.?</td>
<td>1. Yes 2. No -99. Refused to answer</td>
</tr>
<tr>
<td>I do not participate in making major household purchases...</td>
<td>1. Because I don’t like doing/ can’t/ don’t want to do it 2. Because it is not my responsibility 3. Because I don’t think it is important for me to make major household purchases 4. Because others don’t expect/ want me to make major household purchases 5. Because I will get in trouble if I do/ am not allowed to -96. Other (specify): -98. Don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>J.11. Thinking about your role in the household and the expectations of you, in the past 2 weeks have you ever chosen for yourself not to perform any of your roles or responsibilities?</td>
<td>1. Never 2. One or two times 3. Several times 4. Often 5. Nearly every day -98. Not sure/don’t know -99. Refused to answer</td>
</tr>
<tr>
<td>J.13. This is a ten step ladder, where on the bottom, the first step, stand people who are completely coerced or powerless, and on the highest step, the tenth step, stand those with the most ability to advance goals that they value in their own homes and in the world. On which step are you today?</td>
<td>Response Options: 1-10 -98. Don’t know -99. Refused to answer</td>
</tr>
</tbody>
</table>

Notes: Table lists the 64 closed-ended questions used as the set from which the best 5 were selected.