Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 10046

Qualitative Analysis at Scale

An Application to Aspirations in Cox's Bazaar, Bangladesh

Julian Ashwin
Vijayendra Rao
Monica Biradavolu
Arshia Haque
Afsana Khan
Nandini Krishnan
Peer Nagy

WORLD BANK GROUP
Development Economics
Development Research Group
May 2022

Policy Research Working Paper 10046

Abstract

Qualitative work has found limited use in economics largely because it is difficult to analyze at scale due to the careful reading of text and human coding it requires. This paper presents a framework with which to extend a small set of hand-coding to a much larger set of documents using natural language processing and thus to analyze qualitative data at scale. The paper shows how to assess the robustness and reliability of this approach and demonstrates that it can allow the identification of meaningful patterns in the data that the original hand-coded sample is too small to identify. The approach is applied to data collected among Rohingya refugees and their Bangladeshi hosts in Cox's Bazaar, Bangladesh, to build on work in anthropology and philosophy that distinguishes between ambition–specific goals, aspiration–transforming values, and navigational capacity, which is the ability to achieve ambitions and aspirations. The findings demonstrate that these distinctions can have important policy implications.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at vrao@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Qualitative Analysis at Scale: An Application to Aspirations in Cox's Bazaar, Bangladesh*

Julian Ashwin[1,2], Vijayendra Rao[3], Monica Biradavolu[4], Arshia Haque[3], Afsana Khan[3], Nandini Krishnan[3] & Peer Nagy[2]

[1]London Business School, [2]University of Oxford, [3]World Bank Group, [4]Qual Analytics

*The authors would like to thank participants at the Methods and Measurement Conference 2021 and the World Bank's "Half-Baked" seminar and Ikechi Okorie for their useful comments and feedback. Aditya Chhabra and Eleni Kalamara provided valuable research assistance for the project. The authors are grateful to the World Bank's Knowledge for Change Program, and the World Bank-UNHCR Joint Data Center on Forced Displacement for financial support.

1 Introduction

Many questions of interest to economics are better captured with open-ended qualitative interviews rather than structured questionnaires. These include important concepts like well-being, vulnerability, resilience and – the focus of this paper – aspirations. Quantitative, structured questions have become the standard format for questionnaires largely because they are relatively easy to administer to large representative samples of respondents, and generate quantitative data that are easier to analyze using standard econometric methods. Structured questions work best on concepts where the possible range of responses can be correctly predicted in advance by the researcher, and where any needed probes can also be predicted in advance. They also require that respondents have the same understanding of the latent construct underlying the question as the researcher. Thus, the researcher must know with a high probability not only the full range of responses that the question is likely to elicit, but also the types of follow up questions that might be required to delve deeper into an issue.

For these reasons, structured questions do not work well for more complex concepts where respondents have a heterogeneous understanding of the concept, where responses can be difficult to predict, and where probes, and the range of responses, cannot be anticipated in advance. When such concepts are studied with structured questions that force respondents to choose from a limited range of numerical responses, they can result in metrics that have the appearance of being clearly defined but hide the complexity of the "truth", and result in imperfect science and policy (Espeland and Stevens, 1998). Such subtle latent constructs are, arguably, better studied with open-ended methods where the respondent is allowed the full freedom to respond in an open-ended conversational style and in the manner of their choosing, and for the interviewer to probe by interactively reacting to responses.

Open-ended approaches to interviews have not been employed much by economists because analyzing them is hard and almost impossible to do at scale (Rao, 2022). They are primarily the domain of qualitative researchers in anthropology, sociology and related fields who mull over recordings or transcripts of interviews for considerable periods of time, listening and carefully coding them to make sense of them and to discern patterns in the data. It is a very labor-intensive process, and consequently is generally associated with small sample studies that are not statistically representative. The small sample challenge that has, so far, been intrinsic to qualitative methods has resulted in a large methodological literature on qualitative and case-study analysis which focusses on interpreting data from interviews gathered from samples that are not designed to be statistically representative of larger populations. The general approach to date has been to inductively draw out inferences that reflexively expand our understanding of an issue, or to inform theory rather than claim statistical representativeness (Small, 2009).

This paper outlines a new method to analyze open-ended interviews at scale. It shows how open-ended interviews from large-N, statistically representative, samples can be analyzed by combining human coding and machine learning. The method attempts to follow the logic of traditional qualitative analysis as closely as possible. Briefly, a sub-sample of the transcripts of open-ended interviews are coded by a small team of coders who read the transcripts, decide on a "coding-tree" and then code the transcripts using Atlas-TI software which is designed for this purpose. This human coded sub-sample is then used as a training

set to predict the codes on the full, statistically representative sample. The annotated data on the "enhanced" sample are then analyzed using standard regression analysis.

This method has several advantages over "unsupervised" natural language process methods used for analyzing text such as topic modeling (which searches for words that occur in clusters in the data) in that it attempts to hew as closely as possible to traditional qualitative analysis by using the judgement of informed human coders to be scaled-up, rather than have computers make sense of the data. It also has an advantage over methods such as sentiment analysis which maps text against pre-coded dictionaries; sentiment analysis can only provide broad assessments of the "sentiments" observed in the data and is not good for nuanced analysis, and dictionaries in non-European languages are not well developed. Working with human codes in a sub-set of the data falls in the category of "supervised" natural language processing methods – but gives us a training set that is specific to the sample being analyzed, and thus has the potential for nuanced, context-specific analysis. It is thus analogous to a dictionary created specifically for the analytic sample. We apply the method to study parents' aspirations for their children by analyzing data from open-ended interviews conducted on a sample of 1,000 Rohingya refuges and their Bangladeshi hosts in Cox's Bazaar, Bangladesh.

The paper by proceeds next by providing a brief overview of the literature on narrative analysis in economics and aspirations in development economics. We explain how open-ended interviews may add to our understanding of this important topic by opening hitherto understudied dimensions that have important implications for research and policy. Section 3 provides some context to the data - on Cox's Bazaar and the process by which the open-ended interviews were conducted and transcribed. Section 4 explains the human coding process – the development of the coding tree, the process of coding validation and checking, and inter-coder reliability. We then move to the NLP methodology for extending the human coded sample in Section 5. Section 6 sets out and discusses results for both the human and enhanced samples. Finally, Section 7 concludes and makes suggestions for further work.

2 Related Literature

2.1 Narrative Analysis in Economics

The difficulties with using qualitative methods such as discourse analysis at scale on representative samples have led to their largely being neglected in modern economics. There are notable exceptions, such as the widely used monetary policy shock series developed by Romer and Romer (2004) that uses detailed readings of central bank minutes and the narrative approach to business cycles proposed by Shiller (2020). However, the introduction of natural language processing (NLP) methods has led to a recent focus on using text data in a quantitative manner as an important source of information in economic research (Gentzkow et al., 2019).[1]

Most work in economics that uses text in a quantitative way falls into two categories that, while relevant in our context, are conceptually quite different from the method we propose.

[1]This trend is also present in other social sciences, see Ferguson-Cradler (2021) for a discussion of the use of computational text analysis to identify narratives in economic history.

The use of unsupervised statistical models to reduce the dimensionality of text documents into a smaller set of interpretable variables that can then be used in further analysis; and the use of dictionary methods to extract a signal of interest from documents. A recent example of the former approach in development economics is Parthasarathy et al. (2019), who use a structural topic model (Roberts et al., 2016) to decompose the transcripts of village assemblies in rural India. Other examples of such work in non-development contexts includes Hansen et al. (2018), Nimark and Pitschner (2019) and Larsen et al. (2021).

Dictionary methods are common, particularly for the analysis of sentiment, where a wide range of general purpose and context-specific word lists are available. An early example of this is Tetlock (2007) who uses the general purpose General Inquirer's Harvard IV-4 psychosocial dictionary to quantitatively measure interactions between media sentiment and the stock market. However, many economic researchers have proposed context-specific dictionaries that help them extract their particular signal of interest. Loughran and McDonald (2011) introduce a dictionary that classifies words as positive or negative in the context of economic news. These dictionary methods are not limited to the analysis of positive vs negative "sentiment", but have also been developed to measure a wide variety of other information. For example, by Apel and Grimaldi (2012) to measure the "hawkishness" of central bank communication, by Correa et al. (2017) to measure financial stability and Nyman et al. (2021) to measure systemic risk. The influential economic policy uncertainty index developed by Baker et al. (2016) is also based on a simple dictionary based method. The context-specificity of these word lists is of course a limitation as well as an advantage.[2] They are also limited in that they impose a structure on the text features that they try to extract - the presence or absence of certain sets of words.

Our approach is to extend a small set of human annotations conducted by qualitative researchers to a larger representative sample using a model trained on this subsample. We are therefore perhaps closest to literature that combines both qualitative work with NLP methods.[3] Michalopoulos and Xue (2021) use an archive of manually coded motifs in folklore introduced by Berezkin (2015) and then use NLP to classify these motifs into different concepts. A recent paper by Jayachandran et al. (2021) is similar to ours in spirit, as they use a subset of manually coded documents in order to identify which quantitative survey questions best capture women's agency. Although their approach is methodologically quite different the aim is similar: to find a way to scale up the measurement of nuanced and complex concepts to large samples. In ongoing work Alexander et al. (2017) conduct a "qualitative census" of poverty in the United States through open-ended interviews with a representative sample of poor households, which would be a potential use case of the methodology we discuss here. Methodologically, our approach is similar to that of (Mann and Püttmann, 2018) who use a supervised NLP model to extend a sub-sample of patents that have been manually classified as automation and non-automation to a representative sample using a supervised model based on the patents' text.

There is also a related literature outside economics on training supervised models on hu-

[2] In fact, Saiz et al. (2021) suggest that, particularly in a forecasting context, as tailored dictionaries have been constructed with previously observed events in mind, they do not capture unexpected phenomena as well as general purpose methods.

[3] It is quite common to use a subset of manually classified articles to validate a measure derived from text, e.g. Baker et al. (2016) and Shapiro et al. (2020), but our focus is on using the manual classifications to construct a measure.

man annotations. However, while our focus is on whether and how such methods can assist substantive economic analysis, this typically focuses on either maximizing predictive performance or assisting an ongoing coding process.[4] Yordanova et al. (2019) provide a good summary of the literature that focuses on predictive performance. Their paper is quite similar to ours in approach to the NLP modeling, but as their focus is purely on prediction, they include contextual information (e.g. time of day) in the predictive model and do not perform any statistical analysis on the extended sample to show why the use of NLP is helpful here.

Good examples of using NLP to assist the process of human annotation are Liew et al. (2014) and Wiedemann (2019) who propose an "active learning" approach in which a model is trained on a small annotated sample to maximizing the true positives, which are then corrected manually. Meanwhile, Karamshuk et al. (2017) use a hybrid approach where they first get a small number of high quality annotations, and then use these to crowdsource a much larger one and train a neural network on this larger sample. While we think this is potentially a very useful approach, the use of crowdsourced annotations may not be ideally suited to nuanced and complex concepts. Other work, such as Chen et al. (2018), focuses on ambiguity and disagreement across coders, this is certainly an important issue in qualitative work and one where NLP techniques may prove useful, but not the focus of our paper.

2.2 Aspiration, Ambition and Navigational Capacity

There is a thriving literature on aspirations in development economics that emerged from Debraj Ray's seminal paper (Ray, 2006) which extended conventional economic models of human capital investments by arguing that preferences are not exogenously determined but are social - shaped by what an individual observes around in their "cognitive neighborhood" that results in an "aspirations window." This aspirations window can be multidimensional and include things ranging from education and income to dignity and good-health. This idea was then extended by Genicot and Ray (2017) and others, reviewed in Genicot and Ray (2020), to show that socially determined aspirations can fundamentally affect issues that range from education and mobility to collective action and conflict. The development of theory has gone in parallel with a thriving empirical literature, Fruttero et al. (2021) analyze how socially conditioned aspirations matter in a variety of important spheres, and particularly in educational and labor market investments.

The empirical literature is based on quantitative measures of aspirations using structured questionnaires and, perhaps consequently, does not delve into broader dimensions of aspirations that Ray first talked about such as dignity or cultural heritage which are more difficult to measure. It also misses an important point first made by the anthropologist Appadurai et al. (2004) that aspirations are affected not just by an individual's ability to imagine a different future for themselves or their children, and by the economic resources that they can draw on by, but also by their "capacity to aspire" which is a cultural and cognitive resource that allows then to navigate their way to a better future. Furthermore, more recently, the philosopher Agnes Callard has argued that it is important to distinguish between what she calls "ambition" and "aspiration" (Callard, 2018). She defines an "aspiration" as a process

[4]Furthermore, the text features dealt with here are often quite straightforward, so potentially quite different from concepts like aspiration and navigational capacity. In fact, Crowston et al. (2010) find that simple rule-based algorithms perform better for many of their text features than their supervised models.

of reversing a "core value" that results in a "change in the self." An "ambition" to her is a specific goal that which "she is fully capable of grasping in advance of achieving it" (Callard, page 229). Ambition, to her, "often directed at those goods – wealth, power, fame – that can be well appreciated even by those who do not have them." By Callard's definition, the economist's understanding of aspiration is more in line with what she would call "ambition" rather than "aspiration", a definitional distinction that we adopt in this paper as well.

These distinctions are not just semantic. They have implications for measurement. Navigational capacity, being a cognitive and culturally determined capacity, is likely to be less amenable to structured questions where responses to questions are not easy to predict in advance. Similarly, aspirations in Callard's sense, as transformative processes are potentially very differently conceived by different individuals and thus have heterogenous understandings of the latent concept – which also make them difficult to study with structured questionnaires.

These distinctions could also have potentially important implications for policy – if navigational capacity matters it could suggest that interventions to improve cognitive ability might matter, as might interventions to guide less advantaged people towards achieving their goals. If aspirations matter in a way that is different from ambition, then it might be important to distinguish between them in understanding how people might invest time and resources in achieving aspirations vs ambitions, and – perhaps - in designing interventions that, for instance, are delivered by cultural or faith-based institutions rather than by government.

3 Context and Data

The data analyzed in this paper is from Cox's Bazaar in Bangladesh where about 750,000 Rohingya refugees who were forcibly displaced from Myanmar between 2017-18 are primarily housed. The challenges faced by displaced populations and hosting communities go beyond monetary or monetizable welfare measures such as food consumption and security, household expenditures, labor market outcomes and earnings, and basic living standards. Particularly in contexts of forced displacement outside the country of origin, the displacement experience is often accompanied by reliance on humanitarian assistance, lack of documentation, limited or no access to labor markets and services, and limited mobility, at least in the short term.

Host communities at the same time, face a sudden influx of population, increasing pressure on scarce local resources – land, jobs, services for instance, fears of insecurity and illicit activities, and risks to the social cohesiveness of their communities. To the extent that displaced populations move into poorer or lagging hosting areas, with limited capacity to adjust, these pressures may exacerbate pre-existing challenges to welfare and socio-economic mobility among the host community.

The 2017 influx of the Rohingya from Myanmar to Bangladesh has remained overwhelmingly concentrated in the border district of Cox's Bazaar. It has implied a massive increase in localized density in the two primary hosting sub-districts of Teknaf and Ukhia, which were already lagging compared to the rest of the district in terms of human capital, access to services and jobs in growing sectors, and reliance on low productivity agriculture and

service sector jobs. While humanitarian assistance has been largely successful in meeting the basic needs of the displaced Rohingya in terms of food, shelter and water, sanitation and hygiene, like many other forcibly displaced populations, they continue to face challenges in terms of access to formal education for their children, restrictions on freedom of movement and limited livelihood options.

In this context, there is a need to understand well-being beyond what might be measurable in standard surveys. Given the heavy reliance of the Rohingya population on humanitarian assistance to meet basic food needs as well as the limited ability of this population to engage in market-based activities, standard measures of monetary welfare may provide an incomplete assessment of their welfare. At the same time, household survey data may not adequately capture aspects of host community welfare, particularly those that are unobservable and hard to measure using standard quantitative instruments. These include aspects related to the Rohingya influx, such as emerging risks for social cohesion, and unrelated to the influx, such as whether and how a population from a region that is largely disconnected from national growth opportunities, can plan and aspire for upward economic and social mobility – which is the focus of this paper.

This open-ended survey builds on an existing panel survey of about 5,000 randomly selected households from the Cox's Bazaar population split more or less evenly between Rohingya and their Bangladeshi hosts. The baseline survey was conducted between April and August 2019 (World Bank, 2019) and the qualitative, open-ended, questions were conducted in a subsequent survey round implemented between October and December 2020.[5]

The baseline survey administered a 2-part instrument consisting of:

1. A household questionnaire, primarily administered to an adult member of the household (age $>$ 15) who is knowledgeable about the household's day-to-day activities. The household questionnaire included modules on household roster and composition, housing characteristics, food security, consumption, household income, sources of assistance, assets and anthropometrics for children under 5.

2. An adult questionnaire administered to two randomly selected adult members of the household (age $>$ 15) about their individual information and experiences. This included modules on labor market and labor market history, history of migration, access to health services, crime and conflict and mental health.

For the qualitative interviews we attempted to obtain information from a random sample of 25% of the full sample i.e., 1,255 households. Some households we contacted were deemed ineligible because they did not have any children, other households refused to be interviewed, and some of the recordings were inaudible because of phone network disruptions. With this we have a completed sample of 1,040 interviews.

Due to the experimental nature of the open-ended data collection process we decided to restrict the interviews to a maximum of fifteen minutes, and after pre-testing various alternatives settled on the following interview protocol. We began with a short quantitative,

[5]The time lag between the baseline quantitative survey and the open-ended interviews in 2020 clearly represents a challenge since conditions faced by households may have changed in that time period. We therefore use the baseline data largely as controls in reduced form regressions.

structured questionnaire to elicit the households' educational ambitions for their children, which included a few questions on the impact that COVID had on children's education. The qualitative interview protocol followed at the end of this short education module with the following questions:

1. [**Screening question**] What is the name of your eldest unmarried child living in this household?
2. Can you tell me about the hopes and dreams you have for your children?
3. What have you done to help them achieve these goals?
4. What do you plan to do in the future to help them achieve these goals?
5. Can you tell me about the hopes and dreams you have for [**name of eldest child**]?

The qualitative data were collected by 5 interviewers, supervised jointly by a local survey firm and a subset of the authors of this paper. A three-day virtual session was conducted in Bengali to orient the team on the module and train them on conducting open-ended qualitative interviews. The open-ended module was very different from what the interviewers had previously been trained on and thus required higher involvement and substantial handholding throughout the data collection process. The interviewers participated in 4 debriefing sessions, the purpose of which was to help interviewers brainstorm with the full team on appropriate interview techniques and best practices: probing when appropriate, adapting questions according to the respondent's context, learning to listen, being flexible and responding to any ethical challenges. Interviewers of both hosts and refugees were required to be fluent in the local Chittagongian dialect (which is very close to the Rohingya dialect) and interviewers for refugees had an additional requirement: being based in Cox's Bazaar. All the interviews were recorded (with the permission of the respondent), transcribed, and translated into standard Bengali. The transcribed interviews were then merged with the baseline survey data.

The interviews were conducted in Bengali, but we work with English translations of the transcripts. While human translations are available, these are often somewhat inaccurate and often contain spelling mistakes which makes machine analysis difficult. We therefore work with machine translations (using Google Translate) in the analysis presented here. These machine translations are of course also imperfect, but we find that they are less noisy and yield slightly more interpretable results than human translation (the difference is very small however).

In addition to the open-ended interviews, we also use several quantitative variables from the baseline survey on household characteristics. Table 1 shows summary statistics for these variables. In addition to the open ended interviews, subjects were asked to rate their education ambitions on a scale from 1 to 7, where 1 is no education and 7 is postgraduate level. This variable is also used as a measure of education ambition in Section 6 and is summarized in the final row of Table 1.

Table 1: Quantitative variable summary statistics

Statistic	N	Mean	St. Dev.	Notes
Refugee status	1,040	0.450	0.498	Dummy variable, 1 for refugee
Eldest child's sex	927	0.467	0.499	Dummy variable, 1 for female
Eldest child's age	928	10.861	4.963	Integer
Female household head	1,039	0.160	0.367	Dummy variable
Number of children	938	2.617	1.437	Integer
Parent's years of education	1,040	3.544	3.854	Integer
Parent's religious education	1,038	0.041	0.199	Dummy variable
Asset Index	948	0.129	1.862	Principle Component of assets owned following Filmer and Pritchett (2001)
Household Income	948	1.137	2.449	Income for last month in 10,000s Bangladeshi taka
Trauma Event Score	948	8.691	2.705	Sum of positive responses for experience of twelve possible traumatic events following Harvard Trauma Questionnaire
Quant Education Ambition	935	3.890	1.696	Ordered categorical (1-7) from question on parents' ambitions for eldest child's education

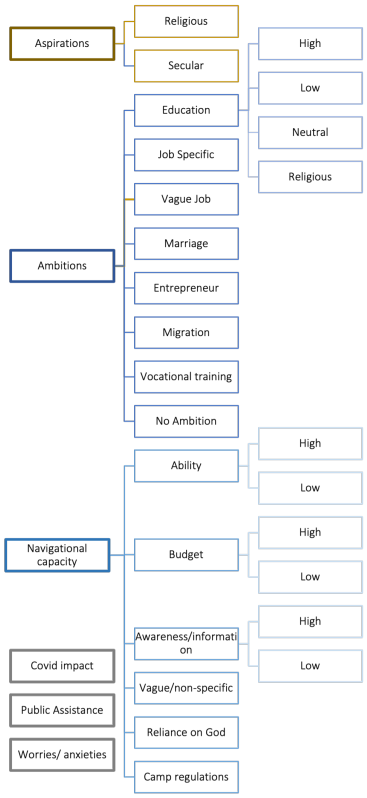
4 Human Qualitative Coding

4.1 Coding Tree

An initial set of 40 transcripts, 10 each, were read by four co-authors (Arshia Haque [henceforth, AH], Afsana Khan [AK], Monica Biradavolu [MB] and Vijayendra Rao) to develop a "coding tree." We ensured that this initial reading included transcripts of male and female, and host and refugee respondents. Detailed discussions led to the development of a coding tree with 3 primary codes and 21 subcodes and associated definitions. Using Atlas-TI, a qualitative data analysis software, a two-person team (AH and AK who were both also responsible for the data collection) coded 395 transcripts, comprising 50% host and 50% refugee interviews which were randomly drawn from the 1,040 transcripts.

Figure 1 shows the coding tree. Aspirations, following Callard's definition, were divided into "Religious" and "Secular" and assigned dummy variables for whether the household expressed them or not. Ambition was divided into six major categories – education (further sub-coded into High, Low, Neutral and Religious), Job Specific, Vague Job (where the job aspiration was not clearly expressed), Marriage, Entrepreneurship, Migration, Vocational Training, and No Ambition. Navigational Capacity was coded into six sub-codes – Low and High "Ability" and Low and High "Budget". There were also three additional codes that did not fit into the structure of aspiration, ambition and navigational capacity, so are not shown in Figure 1 and we will not discuss them further in this paper. These excluded codes were Covid Impacts, Public Assistance and Worries/Anxieties, and we leave an analysis of these dimensions for further work.

Figure 1: Coding Tree



Descriptions and examples of these codes are displayed in Appendix A, but Figure 2 includes a few examples to illustrate some differences between aspirations, ambition and capacity.

Figure 2: Examples of qualitative codes

(a) Ambition:Education:Low

"God willing, I will teach my son up to 10th class. If he wants to stay in Bangladesh for 20-25 years, I want him to get a job here"

(b) Ambition:Education:High

"My daughter's dream is to study. I'll do it. If Allah keeps me alive, I will educate my daughter so she can get a job in administration."

(c) Navigational Capacity:Ability:Low

"I don't do much at home. I help her as much as I can."

(d) Navigational Capacity:Ability:High

"The school is still closed for Corona. So, by selling some of my food, I have arranged for private teacher by paying at minimum."

(e) Aspiration:Secular

"They will become well behaved, good human beings. Will have a respectable job."

(f) Aspiration:Religious

"I don't want make my son work. I want him to become a religious cleric (hujur).."

Atlas-TI software was used to set up the human coded database. The data was first orga-

nized following Atlas-TI's manual on 'column control via field name prefixes' to name each of the documents using their Case IDs as well as to group the documents into preferable segments before using 'survey import' into Atlas Desktop. The project was then set up on the cloud version of Atlas-TI where both coders could work independently. To review coded excerpts, projects were imported back into Atlas' desktop version to generate an Excel spreadsheet with desired variables and quotation sheets segregated by codes.

To achieve agreement between coders, two coders [AH and AK] first applied the codes to 30 transcripts in Atlas-TI. The coded excerpts were shared in an Excel matrix that was reviewed by MB. Any unclear applications of codes were identified, discussed, and resolved in weekly meetings. The process of review and resolution was conducted throughout the coding process, in batches of approximately 60 until all 395 were coded. The continuous review process not only reduced disagreement between coders but also led to the creation of new codes and a deeper understanding, and sharper definitions, of certain codes.

Table 2 illustrates the process by which codes were refined to be more nuanced and context-specific as a result of the review process, for the example of expressions of religious aspirations and ambitions. Initially, when a parent stated that they wanted their child to be a maulvi or be alem/alemdar or hafez, or wanted their child to go to madrassa or noorani school, these instances were coded as Aspiration:Religious. After reviews and seeking expert input, we understood that these references must not just be coded for religious aspiration, but also for religious ambition, specifically for Ambition:Education:Religious. Further, this religious education ambition could be scaled using a Ambition:Education:High, Ambition:Education:Neutral or Ambition:Education:Low code. As a result, the definitions for both the aspirations and the ambition group of codes were better specified, leading to a deeper and finer-grained understanding of respondents' hopes and dreams for their children.

Table 2: Coding religious education

Statement	**Code applied**
Wants child to be a Maulvi/alem	Aspiration:Religious + Ambition:Education Religious + Ambition:Education:High
Wants child to go to madrassa	Aspiration:Religious + Ambition:Education:Religious + Ambition:Education:Neutral
Wants to send child to noorani madrassa	Aspiration:Religious + Ambition:Education:Religious + Ambition:Education:Low
Wants child to be a hafez	Aspiration:Religious

4.2 Disagreement

To account for instances where the two coders (AK and AH) and the coding reviewer (MB) did not agree on a code, we created a 3-level ranking system for each code - "fuzzy", "reliable", and "very reliable". At the end of each batch of coding, the two coders ranked each code on whether they considered their own application of codes to be fuzzy, reliable, or very reliable. The reviewer similarly ranked each code using the same scale. Whenever there was a mismatch in ranks provided by these three individuals, quotations under the said code would be thoroughly refined to reach a clearer definition.

In the example shown in Table 3, MB rated the code "Job Secular" as fuzzy as she ob-

served religious jobs such as "madrassa teacher" was coded under secular by both coders. This was resolved by further refining the "Job Secular" code and creating further subcodes to separate different types of jobs that parents aspired for their children. On the other hand, the "Vocational Training" code considered as "very reliable" because each coder evaluated that the application of this code was unproblematic, and the reviewer agreed with this assessment.

Table 3: Resolving disagreement

Code**Code**	**Description**	**AH**	**AK**	**MB**
Job Secular	Coded when specific job, occupation or work type was highlighted.	Reliable	Reliable	Fuzzy
Vocational training	Any vocational training in the context of ambition is mentioned.	Very Reliable	Very Reliable	Very Reliable

The goal of the process was to ensure that at the end of each review process, both the coders and the reviewer agreed that all codes were assigned the rank of "very reliable".

4.3 Consolidating codes for NLP models

In order to prepare the human annotations to be used as training data for the NLP models we take two steps. First, we aggregate individual annotations to the question-answer (QA) level. Second, we consolidate the 21 sub-codes into 13 higher level codes that are less sparse and so allow for interpretable aggregation to the interview level while still being rich enough for our models to have good predictive performance.

Aggregating individual annotations to the QA level allows us to maintain granularity within interviews in a way that can be straightforwardly extended to the unannotated interviews. As the interviews have a flexible back-and-forth question-answer format, each QA pair generally has a single focus (the average question is 12.6 words long and the average answer is 20.3 words). As the annotations themselves regularly span multiple sentences and answers often do not make sense without the context of the question, it is preferable to work at the QA level rather than at the sentence level. The 1,050 interviews are thus split into 8,119 distinct QA pairs. Of these, 3,202 are in the human annotated sample and 4,917 are unannotated. Each annotated QA pair is potentially associated with any number of the sub-codes: 2,090 of the 3,202 have at least one annotation and pairs have 1.1 codes associated with them on average.

Consolidating the 21 sub-codes into a smaller number of more general codes allows us to prevent any individual sub-code from being so sparse that our (cross-validated) NLP models do not have good performance. It also has the added benefit of allowing us to construct quasi-continuous variables for ambition and navigational capacity at the interview level. The consolidated codes are shown in Table 4. Marriage and Migration are kept as individual codes as they do not straightforwardly fit into high or low ambition and feature in interviews sufficiently frequently to allow good model performance. Camp Regulations are also kept as individual codes as it will largely apply to refugees and so would bias capacity scores of refugees if included as Low Capacity. These consolidated codes are aggregated at the QA level by taking the maximum of their associated sub-codes, giving a binary variable

for each code.

Table 4: Consolidated Sub-Codes

Code	**Sub-Code**
Secular Aspiration	Aspiration:Secular
Religious Aspiration	Aspiration:Religious
Religious Education	Ambition:Education:Religious
Low General Ambition	Ambition:Education:Low, Ambition:No_Ambition, Ambition:Education:Neutral, Ambition:Entrepreneur, Ambition:Vocational_Training, Ambition:Vague_Job
High General Ambition	Ambition:Education:High, Ambition:Job_Secular
Low Education Ambition	Ambition:Education:Low, Ambition:Education:Neutral
High Education Ambition	Ambition:Education:High
Low Capacity	Capacity:Reliance_on_God, Capacity:Awareness/Info:Low, Capacity:Ability:Low, Capacity:Vague/Non-Specific
High Capacity	Capacity:Ability:High, Capacity:Awareness/Info:High
Budget Low	Capacity:Budget:Low
Budget High	Capacity:Budget:High
Marriage	Ambition:Marriage
Migration	Ambition:Migration
Camp Regulations	Capacity:Camp_Regulations

In particular, consolidating the codes for navigational capacity into a high-low scale allows us to incorporate a variety of concepts including Ability, Informational Awareness and Reliance on God into a single quasi-continuous variable.

Given the QA-level binary codes we can then create quasi-continuous variables at the interview level. The Aspiration, Religious Eduation, Marriage, Migration and Camp Regulations codes are simply a mean of the binary classifications at the QA-level. So if an interview has a Religious Aspiration score of 0.25 this means that 25% of the QA pairs in that interview have been coded as Religious Aspiration. For the ambition, capacity and budget codes, we create a variable on a low to high scale, by classifying pairs with a "low" code as 0, those with a "high" code as 1 and dropping pairs with neither a high nor low code. For example if an interview has three QA pairs coded as Low Capacity and one coded as High Capacity it would have an interview-level capacity score of 0.25.[6] Summary statistics for these codes on the human annotations only are shown in Table 5.

[6]Other methods of aggregation are possible here but, as our NLP models are trained at the QA pair level with aggregation performed afterwards, as discussed in Section 5, this will not affect our method of extending the human annotated sample.

Table 5: Summary Statistics of Human Annotations

Statistic	N	Mean	St. Dev.
Secular Aspiration	395	0.065	0.121
Religious Aspiration	395	0.037	0.082
Religious Education	395	0.037	0.083
General Ambition	387	0.297	0.329
Education Ambition	355	0.247	0.340
Capacity	340	0.409	0.391
Budget	257	0.252	0.404
Marriage	395	0.088	0.145
Migration	395	0.023	0.085
Camp Regulations	395	0.017	0.061

Note: $N \neq 395$ for the variables on a high-low scale as if there are neither any low nor high codes for any of the QA pairs of that interview, the variable will not be defined.

5 Modeling approach

In this Section, we describe the NLP modeling approach we use to scale up our small sample of human annotations to the whole corpus of interviews. First, we describe in general terms how our strategy of enhancing a human coded sample with NLP works. Second, we provide some greater detail on the supervised models, text representations and training method we use. We then set out how to adjust for the added measurement error around the machine annotations relative to the human annotations both by testing for bias in these errors and accounting for the additional variance. We then illustrate our method with a semi-synthetic example and finally discuss validating our supervised NLP models.

5.1 Creating an enhanced sample

For a total of N interviews, let N_h be the number for which we have high-quality human annotations and $N_m = N - N_h$ the number of interviews which have not been human annotated. Our goal is to to create an "enhanced" sample in which we retain the N_h human annotations but add machine annotations for the remaining N_m interviews.

As the qualitative codes are defined at the level of question-answer (QA) pairs, we train our supervised classifiers at this level. Training at this more granular level, rather than at the level of the whole interview has two advantages. Firstly it allows for our qualitative coders to be more precise in their annotation and potentially pick up multiple contradictory signals within a single interview. Secondly, it gives our NLP models a greater number of more precise observations on which to be trained, while splitting up the interviews in a way that we can replicate in the unannotated sample.[7]

[7]If the data was not in a question-answer interview form, the annotation and training could be done at the sentence level.

As described in Section 4.3, we group the qualitative codes into 14 categories. Each of these then acts as a binary classification on each QA pair. For example, a given QA pair is either coded 1 for high capacity (if the subject *is* displaying high capacity in this interaction) or 0 for high capacity (if the subject *is not* displaying high capacity). This allows us the train a separate supervised classifier for each of the 14 codes. Each model is then focused on identifying only the text features that are associated with that one code.

Once we have trained these models for each code on the human annotated interviews, we then use these models to predict annotations, also at the QA pair level, for the unannotated interviews. Once we have these lower level predictions, we can then aggregate these to the interview level to give predicted scores for each interview subject, as described for the human annotations in Section 4.3. The lower level annotations are aggregated in the same way for the human annotations and machine annotations to give interview level scores. For the variables defined on a high-low scale, there are some interviews in which no QA pairs are tagged as either high or low, so in these cases the ambition/capacity score is undefined. However, this occurs rarely in both the human and machine annotated interviews.

Once we have these interview-level annotations we can create our "enhanced" sample of N observations. We do this by using the N_h human annotations where available and model predictions for the remaining N_m.

5.2 Supervised models and cross-validation

As we are training a separate model for each code, and each code is a binary variable, our prediction problem is simply a series of binary classifications. There are many supervised classification models in the NLP literature that we could use here but, given that we are working with very small samples we found that simpler models generally performed better.

We take a two-stage prediction approach, first creating an unsupervised numerical representation of the text features and then using these features as inputs in a supervised model. Importantly, we use only features of the text as inputs and no data on household characteristics. We tested a variety of text feature representations including term frequency-inverse document frequency (tf-idf) vectors and various pre-trained embeddings, based on either the original transcriptions in Bengali or English translations. In general we found that tf-idf vectors of ngrams (1-3) from the English translation performed well and were fairly robust across the different codes so we present results using these inputs as our baseline case. However, Appendix C compares these baseline results to those with 8 alternative text representations, both in terms of predictive performance and similarity of the resulting enhanced samples. We find that in all cases, the enhanced sample under alternative text representations is very similar to our baseline case, and that no representation consistently performs better than the baseline in terms of out-of-sample prediction error across all codes.[8]

Once we have a numerical representation of the text documents, we train a separate binary classifier for each code. As a baseline case, we use L1-regularized logistic regressions where the regularization parameter and the maximum size of the feature space are estimated

[8]Whether different representations work better for certain contexts is an interesting question worth exploring in further work.

by 3-fold cross validation. As our codes are in some cases very unbalanced, there are many more zeros than ones, we also cross validate the acceptance threshold. In what follows, we will discuss results in cases where we both use the entire available sample as our training set and keep a held-out sample to evaluate the models' predictive performance. In Section 5.5, we describe further the steps we take to validate our models in terms of out-of-sample predictive performance and interpretability.

5.3 Accounting for additional measurement error

In general, the problem of scaling up qualitative work can be thought of as attempting to measure a feature y_i which is encoded in text documents w_i. We have a small number (N_h) of human annotated documents for which qualitative analysis allows us to observe y_i, but for the remaining (N_m) documents we need to generate a predicted annotation $\hat{y}_i = f(w_i)$ based on the text. However, as our predictions are imperfect, we will be introducing additional measurement error. Fortunately, we can to some extent quantify these measurement errors and identify their properties and deal with them accordingly.

5.3.1 Bias

It is important to establish that the measurement error we introduce for the machine annotated documents are not introducing a bias in the measurement of y_i. For example, if we found that measurement errors for ambition where systematically positive for refugees and systematically negative for hosts, we might generate a spurious result purely because of this measurement error.

As we are interested in how aspirations, ambitions and navigational capacity are related to subjects' characteristics, as introduced in Section 3, we can test whether the prediction errors for those observations where we have human annotations are related to these 10 characteristics. We do this by regressing the prediction errors from our cross-validated model on these characteristics and computing the F statistics of these regressions. Intuitively, if these F statistics are not statistically significant, there is no evidence of bias in the predictions relative to these characteristics.

These F statistics for each of the codes are shown in Table 6.[9] We see that only for the Camp Regulation code is this significant at the 5% level, so for the other codes a bias introduced by the measurement error in the machine annotations need not be a major concern. The fact that the Camp Regulation code is potentially problematic is perhaps unsurprising as this code appears much more for refugees than for hosts since the hosts do not reside in camps. Our education ambition variable has an F statistic that is significant at the 10% level, which appears to be driven by the fact that prediction errors are more positive for interviews with parents that have more years of formal education. We may therefore wish to treat results on the enhanced sample for Camp Regulation and Education ambition with a degree of skepticism.

[9] The full regression outputs are shown in Appendix C.

Table 6: Testing for biases in NLP model for each code

Code	**Observations**	**F statistic**
Secular Aspiration	311	1.123
Religious Aspiration	311	1.074
Religious Education	311	0.933
General Ambition	298	1.033
Education Ambition	274	1.804*
Capacity	242	0.435
Budget	180	0.819
Marriage	311	1.571
Migration	311	0.681
Camp Regulation	311	1.933**
Note:	*p<0.1; **p<0.05; ***p<0.01	

5.3.2 Variance

Even if measurement error does not introduce a bias in the machine annotations, they will add extra noise to these observations. However, we can quantify the variance of this noise and account for it in our analysis. Following Elbers et al. (2003), we account for two of the types of error in our machine annotations: idiosyncratic error (i.e. the prediction error) and model error (i.e. the sampling errors in the model).[10] Accounting for idiosyncratic and model errors at the interview level is more straightforward than at the question-answer level as it gives us a quasi-continuous dependent variable for which we can assume a normally distributed idiosyncratic error.

To approximate the model error, we bootstrap the model by sampling the interviews without replacement B times, each time we sample a training set of 275 leaving 120 held out observations.[11] This means that our variance estimates are likely overestimates as they are based on training sets that are smaller than the full sample. This gives us an empirical distribution over the predictions based on the sampled distribution. The variance of the machine annotations, taking model error into account, can then be approximated by the variance across all of these bootstrap samples

$$\hat{\sigma}^2_{\bar{y}} = \frac{1}{BN} \sum_{i=1}^{N} \sum_{b=1}^{B} \left(\bar{\hat{y}} - \hat{y}_{b,i} \right)^2 \tag{1}$$

where $\bar{\hat{y}} = \frac{1}{BN} \sum_{i=1}^{N} \sum_{b=1}^{B} \hat{y}_{b,i}$. This can be calculated either in the training set only, or also in the out-of-sample predictions, but we find that the estimates are virtually identical in each.

[10] The authors thank Berk Ozler for his suggestions on this point.

[11] Variances for sampling with replacement are shown in Appendix B. We find that this does not substantially change the variance estimates. If sampling with replacement we ensure that the maximum number of times each interview appears in the training set is less than the number of cross-validation folds.

The idiosyncratic error can then be calculated as the difference between the observed y_i and $\hat{y}_i$. To ensure that these predictions are truly out of sample, we only use observations that were not included in the training set for each bootstrapped run. This variance estimate is also likely a substantial overestimate as it is based on models with a smaller sample size than the full sample. Indeed, using validation set errors to compute this variance leads to substantially lower variance estimates, but we use the higher estimates in the interests of being conservative. The estimated variance of this idiosyncratic error, $\hat{\sigma}^2_\epsilon$ is then the variance of the prediction errors across all bootstraps, for those observations which were not selected into the training set for that run. Of course, this variance has to be calculated on only the human sample, as these are the only observations for which we can observe y_i.

Assuming that these errors are normally distributed, if the idiosyncratic and modeling errors are independent then the estimated variance the machine annotated sample will be the sum of these two variances.

$$\hat{\sigma}^2_m = \hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_\epsilon \tag{2}$$

The estimated variance of the human annotated sample ($\hat{\sigma}^2_h$) is simply the variance of the N_h observed human annotations. This gives us an estimate for the enhanced sample as a weighted sum of the estimated variances for the human and machine annotated samples.

$$\hat{\sigma}^2_{enh} = \frac{N_h\hat{\sigma}^2_h + N_m\hat{\sigma}^2_m}{N} \tag{3}$$

This demonstrates that even if our measurement errors are unbiased there is still potentially a trade-off due to the increase in variance. As our NLP models are imperfect, we would in general expect $\hat{\sigma}^2_m > \hat{\sigma}^2_h$. Thus while enhancing our sample does increase the number of observations it also increases the noise in the sample.

Whether this sample-size vs variance trade-off is worth accepting of course depends on the context in which we intend to use our enhanced sample. However, we can illustrate it with the standard error on an estimate of the population mean. The standard error on the estimated mean using the enhanced sample will include the weighted sum of the variance terms for the human and machine annotated observations.

$$\hat{se}_{enh} = \sqrt{\frac{N_h\hat{\sigma}^2_h + N_m\hat{\sigma}^2_m}{N^2}} \tag{4}$$

The standard error on the estimate for the human sample will be of the usual form. The standard error in the enhanced sample will therefore be smaller than that in the human annotated sample if

$$\sqrt{\frac{N_h\hat{\sigma}^2_h + N_m\hat{\sigma}^2_m}{N^2}} < \frac{\sigma_h}{\sqrt{N_h}} \tag{5}$$

which we can express as a condition on the ratio of the variance in the human and machine annotations.

$$\frac{\hat{\sigma}^2_m}{\hat{\sigma}^2_h} < \frac{N_m + 2N_h}{N_h}$$

Note that the right hand side here will always be greater than one, but the condition shows that adding only a small number of highly noisy machine annotations may not make estimates of the population mean more precise. For our entire sample, where $N_h = 395$ and

$N_m = 645$, then our enhanced sample will have a smaller standard error for an estimate of the population mean if $\frac{\hat{\sigma}^2_m}{\hat{\sigma}^2_h} < 3.6$.

Table 7 shows these variances computed at the interview level for cross-validated models on the full training sample and the standard errors for the population mean that have been adjusted as described above. We can see that in all cases the standard error of the population mean is lower than that of the human only sample, suggesting that enhancing the sample with our method will increase the precision of estimates. This is in spite of the fact that, because of the small sample size we are working with, the predictive performance of our models is sometimes relatively low.

Table 7: Measurement error variances and enhanced sample standard errors

Code	N_h	N_m	$\hat{\sigma}^2_h$	$\hat{\sigma}^2_m$	$\hat{\sigma}^2_\epsilon$	$\hat{se}_h$	$\hat{se}_m$	$\hat{se}_{enh}$
Secular Aspiration	395	645	0.015	0.012	0.011	0.006	0.006	0.004
Religious Aspiration	395	645	0.007	0.008	0.004	0.004	0.004	0.003
Religious Education	395	645	0.007	0.007	0.007	0.004	0.005	0.003
General Ambition	387	645	0.108	0.084	0.072	0.017	0.016	0.012
Education Ambition	355	642	0.115	0.085	0.106	0.018	0.017	0.013
Capacity	340	642	0.153	0.153	0.155	0.021	0.022	0.016
Budget	257	588	0.162	0.163	0.101	0.025	0.021	0.017
Marriage	395	645	0.021	0.025	0.003	0.007	0.007	0.005
Migration	395	645	0.007	0.005	0.003	0.004	0.004	0.003
Camp Regulations	395	645	0.004	0.002	0.004	0.003	0.003	0.002

Note: Variances of the measurement errors are calculated by bootstrap with 100 samples, based on sampling 275 documents without replacement from the human annotated documents.

We can thus think of the machine annotated sample as being subject to an additional measurement error due to model and idiosyncratic noise. We can check for biases in these errors and estimate their variance in the manner described above. Once the measurement error has been quantified, we can make the appropriate adjustments depending on the context. If the text-based variables are included in a regression on the left hand side, then we have a classical measurement error so, providing the errors are unbiased, we do not need to make further adjustments. If the text-based are on the right hand side, there will likely be an attenuation error that biases coefficients towards zero.

5.4 Illustrative semi-synthetic example

To illustrate some of the properties of our method and how it depends on the relative sizes of the human and machine annotated samples, we conduct a semi-synthetic simulation exercise. Text is a very high-dimensional and complex input which is difficult to realistically simulate or prove analytical results for, as one might in traditional econometrics. For this

reason, semi-synthetic exercises have become common in the NLP literature.[12] These semi-synthetic exercises typically use real text documents but simulate some numerical variables based on the text. This gives the advantage of realistic text documents, while being able to observe and control the relationship between the text and numerical variables. In our case, we use the interviews as the text documents and then synthetically generate a code for the entire sample that has the same structure as our qualitative codes. This helps us illustrate that expanding qualitative codes to a larger sample with NLP can help elucidate effects that might not otherwise be identified.

To construct the semi-synthetic dataset we estimate a 10 topic LDA model on the interviews at the QA pair level and then chose a topic that is slightly more prevalent among hosts than refugees.[13] We then assign a 1 to QA pairs where the proportion of this topic is greater than 0.12 (so around 10% of answers are coded as 1). We then train our baseline NLP model on these codes, varying the size of the training set, and aggregate these answers to the interview level, giving a quasi-continuous score between 0 and 1, as in the real data exercises described above.

As the codes are synthetically generated for the whole sample, we are able to observe the "true" underlying effect which we can think of as the entire sample being human coded. We can then compare results with an enhanced sample and only a subset used to train the model to this true effect. As we increase the size of our training set, these will of course converge until we use the entire sample as a training set and there is no "enhanced" sample.

Our statistic of interest here is the difference in prevalence between refugees and hosts of the synthetic code. We quantify this with the confidence intervals on a t-test for the difference in means across the two groups. For the human sample this is straightforward, but for the enhanced sample we use the standard error adjustments described in Section 5.3. Let $\hat{\mu}_{h,ref}$ and $\hat{\mu}_{h,host}$ be the estimated mean prevalence of the code in human coded interviews for refugee subjects and host subjects respectively, while $\hat{\mu}_{enh,ref}$ and $\hat{\mu}_{enh,host}$ are the analogous means in the enhanced sample. As the statistic of interest is the difference between the means of the two groups, the standard errors of interest will be

$$se_{h/enh} = \sqrt{\frac{\hat{\sigma}^2_{h/enh,ref}}{n_{h/enh,ref}} + \frac{\hat{\sigma}^2_{h/enh,host}}{n_{h/enh,host}}} \tag{6}$$

Where $\hat{\sigma}^2_{h/enh,ref}$ and $\hat{\sigma}^2_{h/enh,host}$ are the estimated variances of the synthetic code in the human/enhanced sample for the refugee and host populations respectively. These standard errors then allow us to construct a 95% confidence intervals for $\mu_{h/enh,host} - \mu_{h/enh,ref}$ given a human/enhanced sample. We can then compare these confidence intervals to the "true" confidence interval that we calculate using the full sample of synthetic codes.

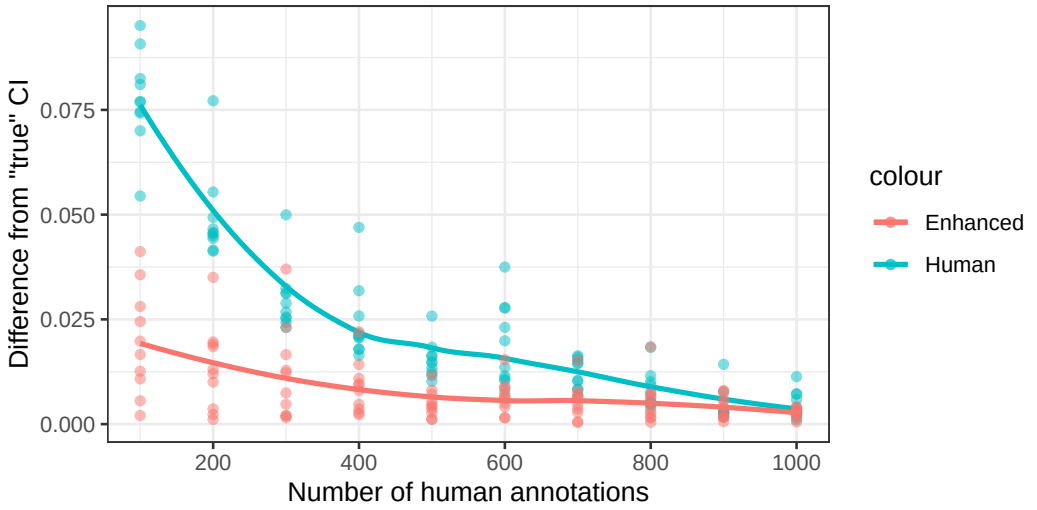
Figure 3 below shows how the distance of human and enhanced confidence intervals from the "true" confidence interval varies with the number of human annotations (i.e. the size of the training set). We see that the enhanced sample substantially reduces the sampling

[12] See Veitch et al. (2020) and Ahrens et al. (2021) as examples from the literature on causal inference with text data.

[13] The LDA model is estimated by Gibbs sampling where the Dirichlet hyperparameters are estimated, with starting values suggested by Griffiths and Steyvers (2004).

uncertainty and does not appear to introduce a bias. To construct this figure, we vary the number of "human annotations" N_h from 100 to 1,000. To ensure that results aren't driven by which interviews happen to be in the human annotated set, we take $S = 10$ separate assignments of interviews into the set of human annotations for each N_h. For each S, estimate the enhanced sample $B = 10$ times by bootstrapping with replacement, to accounts for the model uncertainty in estimated models. For each N_h we then have $S \times B = 100$ enhanced samples. The variation in S accounts for the fact that we could have chosen different interviews for human annotation. The variation in B accounts for the model uncertainty given the draw for S. The variation over N_h accounts for different performance when we vary the number of interviews that have human annotations.

Figure 3: Absolute distance of confidence intervals given enhanced and human samples from ground truth as number of human annotations increases



Note: For each value of N_h (number of human annotations), the blue (red) dots indicate the distance of the human (enhanced) confidence intervals from the true confidence intervals for each of the 10 assignments into human/machine annotated sets.

As we can see in Figure 3, as the size of the human annotated set increases the human confidence interval approaches the truth. This is also true of the enhanced sample: as the number of human annotations increases the NLP model has more data to train on and so will have better predictive performance. This means that both the point estimates will be more accurate and the estimated variance of these estimates will be lower. There is thus likely a trade-off between the cost of additional annotations and the quality of the enhanced sample. Furthermore, it is unsurprising to note that as the size of the human annotated set gets very large, enhancing the sample adds little value. While this trade-off will of course be highly context specific, depending on the nature of the text documents and the concepts being annotated, we believe that enhancing samples with NLP can be useful even in an environment with a very small number of human annotations.

5.5 Validating NLP models

As the aim of our methodology is to yield interpretable results that are trusted by social scientists and policy makers, we take two complimentary approaches to validating our predictive NLP models. First, we assess the validation-set performance of our models to show the use of the text-based models adds information. Second, we use a supervised topic model to illustrate that our models are themselves interpretable.

5.5.1 Validation set performance

As described in Section 5.3, we can quantify the out-of-sample prediction error of our NLP models using held-out observations. This allows us then to quantify how well our models will predict the annotations on the unannotated documents. Although this prediction error is already taken into account in the estimates of measurement error variance, where we demonstrate that even relatively poor predictive performance can lead to usefully enhanced samples provided there is no bias, it may still be useful to take validation-set performance at the QA pair level into account when assessing the reliability of the sample enhancement in a specific context. These validation set performance score provide an additional indication of out-of-sample accuracy.

Table 8 shows the validation set performance for each model estimated at the QA pair level. We display the F1 score (the harmonic mean of precision and recall) for each fold of the validation set as well as the average across folds. Many of our variables are unbalanced to different degrees (in that average number of ones differs across codes) which needs to be taken into account when interpreting the F1 scores. We therefore also add the F1 score of a model that does not use the text at all as a benchmark in the second column. This model randomly assigns codes with a probability chosen to maximize the F1 score, so it can be thought of as a baseline value to indicate whether using the text improves our ability to predict unobserved codes. In all cases, our model substantially outperforms this benchmark, indicating that the NLP models add value for all codes.

Table 8: Validation set F1 scores for the QA pair level models

Code	Benchmark	Fold 1	Fold 2	Fold 3	Mean F1 score
Secular Aspiration	0.092	0.569	0.549	0.518	0.545
Religious Aspiration	0.074	0.642	0.696	0.540	0.626
Religious Education	0.066	0.458	0.494	0.642	0.531
High General Ambition	0.026	0.645	0.672	0.710	0.676
Low General Ambition	0.036	0.663	0.683	0.678	0.674
High Education Ambition	0.064	0.449	0.429	0.542	0.473
Low Education Ambition	0.138	0.633	0.675	0.633	0.647
High Capacity	0.171	0.513	0.458	0.485	0.485
Low Capacity	0.261	0.489	0.464	0.569	0.508
High Budget	0.057	0.627	0.414	0.543	0.528
Low Budget	0.211	0.698	0.661	0.664	0.674
Marriage	0.171	0.814	0.870	0.881	0.855
Migration	0.034	0.667	0.863	0.698	0.743
Camp Regulations	0.036	0.462	0.467	0.353	0.427

5.5.2 Interpretability

The models we use to predict our out-of-sample annotations are not designed to offer great interpretability.[14] While our interpretations themselves come from the definitions of the qualitative codes, we can use models that are more geared to interpretability to analyze our predictions and show that they pick up on relevant features rather than overfitting to noise.

Topic models are often used for descriptive purposes in social science for their interpretability in small samples, see for example Parthasarathy et al. (2019), so are a good choice here.[15] We therefore train a supervised topic model Blei and McAuliffe (2008) on the *predicted* annotations. This model learns topic representations for the interviews which predict the machine annotations. Crucially, we are not using these supervised topic models to enhance the sample, we use it to describe the predictions in the enhanced sample. We estimate a 10 topic model for each of our qualitative variables. For each variable we have 10 topics, defined as a probability weight on each word in the vocabulary. Each of these topics is then associated with a coefficient that tells how likely a document in which that topic is prevalent is to have a high value for that variable.

Figure 4 shows the topics associated with the machine annotations on General Ambition and Capacity. The topics are represented by the terms that appear most frequently in that topic compared to other topics. The coefficient estimates, with a 95% confidence interval, show the effect that topic has on the capacity score for that sentence. We can see that

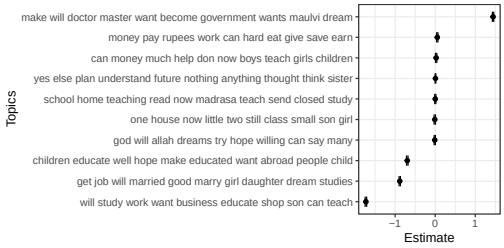
[14] For example, our logistic regressions on tf-idf vectors can have hundreds of coefficients even with the enforced sparsity.

[15] While topic models often give good descriptive results in small samples, their complexity often yields poor out-of-sample performance.

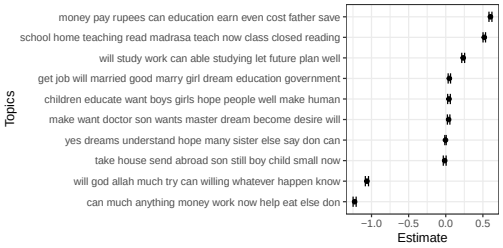
for General Ambition, interviews that spend a lot of time on specific professions such as "doctor", "master" and "government" are likely to score highly, while discussion of marriage and working in a shop or business are likely to score lower. For Capacity, discussion of education and finance increase the likelihood of a high capacity score, while religious language leads to lower scores. Figures for the other qualitative variables are shown in Appendix D. The interpretability of these topic models is still limited by the relatively small sample of predictions, but they are an encouraging indication that the classification of out-of-sample documents matches our interpretations of the codes.

Figure 4: Supervised LDA on model predictions

(a) General Ambition



(b) Capacity



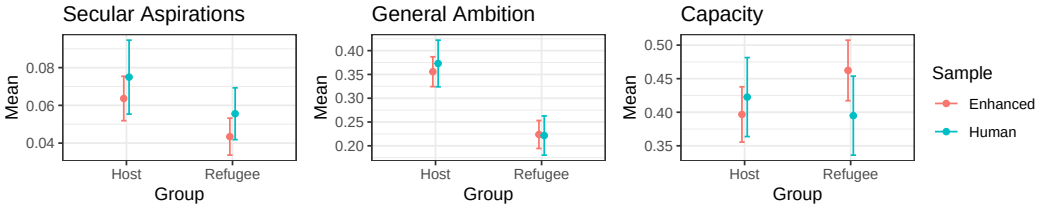
6 Results

The method we have developed in this paper, using Natural Language Processing methods to expand a human-coded subset of open-ended qualitative responses to a more representative "enhanced" sample, can be used to study a variety of questions. Here, we apply this method to explore the broad question of parental aspirations for children in Cox's Bazaar in Bangladesh which houses Rohingya refugees and their Bangladeshi hosts. We analyze whether open-ended responses can elicit information on concepts that are difficult to explore with structured responses – such as distinguishing between "ambition" and "aspiration", using Callard's philosophical distinction between those concepts, and Appadurai's anthropological idea of a "capacity to aspire" – which is the cultural and psychological capacity to navigate one's way towards a set of ambitions or aspirations. While asking questions on these issues, respondents also talked about other issues such as their desire to see their children married, or to have them migrate to another country for a better life.

In this section we analyze the coded data for both the human-coded and enhanced samples. First, we compare the mean value of each qualitative coded variable among host vs refugee subjects, and among parents with a male vs female eldest child. We then regress the codes in OLS reduced form specifications with a set of key variables extracted from the baseline survey conducted in 2019.[16] The explanatory variables, in addition to refugee status and child's sex, include the child's age, whether they belong to a female headed household, the number of children in the family, the respondent parent's years of schooling and whether they have had a religious education, the Filmer-Pritchett asset index, the household's total income in the previous month, and their trauma event score.

[16]Tables here report the analytic standard errors for OLS. Bootstrapping the standard errors leads to broadly similar results but slightly smaller standard errors.

Figure 5: Secular Aspiration, General Ambition and Capacity by Refugee status



Note: Error bars show 95% confidence intervals on estimated means, taking into account measurement errors for the enhanced sample as described in Section 5.3.

An important methodological point to note in our comparison of group-level means is that in all cases the standard error of the difference in means falls in the enhanced sample relative to the human sample. These standard errors include the adjustments for measurement error. We can see an example of this in Figure 5, which shows the refugee and host group means with confidence intervals for Secular Aspiration, General Ambition and Capacity. In all cases, the confidence intervals around the group-level means are tighter in the enhanced sample.

Table 9 compares the host vs refugee and male vs female group-level means for each code and shows the statistical significance of the difference between them.

Table 9: Difference in means for Hosts vs Refugees and Male vs Female

Code	Host - Refugee		Male - Female	
	Human	*Enhanced*	*Human*	*Enhanced*
Secular Aspiration	0.019	0.020***	0.011	0.009
	(0.012)	(0.007)	(0.012)	(0.008)
Religious Aspiration	-0.009	-0.002	0.025***	0.033***
	(0.008)	(0.006)	(0.009)	(0.006)
Religious Education	-0.016**	-0.004	0.016*	0.025***
	(0.008)	(0.006)	(0.009)	(0.006)
General Ambition	0.151***	0.132***	0.039	0.039
	(0.033)	(0.022)	(0.036)	(0.025)
Education Ambition	0.156***	0.091***	0.023	0.046**
	(0.036)	(0.022)	(0.039)	(0.024)
Capacity	0.028	-0.066**	-0.065	0.005
	(0.043)	(0.031)	(0.046)	(0.034)
Budget	0.150***	0.074**	-0.090*	-0.038
	(0.050)	(0.035)	(0.054)	(0.038)
Marriage	-0.024***	-0.019**	-0.078***	-0.109***
	(0.015)	(0.010)	(0.015)	(0.010)
Migration	-0.006	-0.008*	0.026***	0.021***
	(0.009)	(0.005)	(0.007)	(0.005)
Camp Regulations	-0.032***	-0.023***	0.009	0.005
	(0.006)	(0.004)	(0.006)	(0.004)
Note:				*p<0.1; **p<0.05; ***p<0.01.

The previous economics literature has focused on "ambition," which also we study here with codes for "general" ambition (such as a particular job or income level), and "educational" ambition – which is a desire to have a child achieve a given level of education. Table 9 shows that hosts tend to have higher levels of both general and education ambition than refugees. Regression results reported in Table 10 show that these differences are tempered once other explanatory variables are included. Much of the variation in ambition is explained by the parent's years of schooling with more educated parents expressing higher levels of both, general and education, ambition for their children. Note that refugee parents on average have 1.84 years of schooling, while hosts have about 5 years schooling. Parents also tend to express higher ambitions for male over female children.

Ambition is a concept for which structured questions have been developed and used by several scholars. Given that it is a relatively straightforward and easy to understand notion, it is - arguably - better suited for standard survey questions rather than open ended responses. We compare survey responses to education ambitions to education ambition coded from open-ended responses to the question on "hopes and dreams for your child" in the regression reported in Table 10. The standard survey response has much greater explanatory power, and continues to show a strong negative effect for refugees. This could suggest that if a latent construct can be easily translated into an effective structured question, then

the structured response may have less measurement error than an open-ended response to a similar question.

Table 10: Ambition and household characteristics

	Dependent variable:					
	General Ambition		Education Ambition		Quant Education Ambition	
	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)
Refugee status	−0.033	−0.060*	−0.038	0.008	−1.850***	−1.607***
	(0.059)	(0.035)	(0.062)	(0.036)	(0.260)	(0.151)
Eldest child's sex	−0.065*	−0.052**	−0.036	−0.044*	0.085	−0.043
	(0.037)	(0.022)	(0.039)	(0.023)	(0.158)	(0.095)
Eldest child's age	−0.003	−0.006*	0.006	−0.004	0.017	0.022
	(0.005)	(0.003)	(0.005)	(0.003)	(0.022)	(0.013)
Female household head	−0.036	0.031	−0.075	0.039	−0.049	0.056
	(0.052)	(0.031)	(0.056)	(0.032)	(0.224)	(0.131)
Number of children	0.019	−0.004	0.007	0.009	−0.112	−0.177***
	(0.016)	(0.010)	(0.017)	(0.010)	(0.070)	(0.043)
Parent's years of education	0.017***	0.009***	0.019***	0.008**	0.039	0.047***
	(0.006)	(0.003)	(0.006)	(0.003)	(0.024)	(0.014)
Parent's religious education	0.008	0.051	0.041	0.041	0.688	0.577**
	(0.095)	(0.058)	(0.099)	(0.058)	(0.438)	(0.242)
Asset Index	0.015	0.012	0.016	0.021**	0.096	0.065*
	(0.016)	(0.009)	(0.017)	(0.009)	(0.070)	(0.039)
Household Income	0.013	0.003	−0.016	−0.0003	0.147**	0.032*
	(0.016)	(0.005)	(0.018)	(0.005)	(0.066)	(0.019)
Trauma Event Score	0.002	−0.00001	0.004	−0.002	−0.038	−0.001
	(0.007)	(0.004)	(0.008)	(0.005)	(0.032)	(0.019)
Constant	0.227**	0.372***	0.124	0.206***	4.504***	4.613***
	(0.089)	(0.051)	(0.095)	(0.052)	(0.379)	(0.216)
Observations	307	810	286	742	286	774
R^2	0.098	0.060	0.084	0.037	0.452	0.418
Adjusted R^2	0.067	0.048	0.051	0.024	0.432	0.411
Residual Std. Error	0.315	0.312	0.327	0.306	1.308	1.289
F Statistic	3.199***	5.107***	2.532***	2.817***	22.707***	54.847***
Note:						*p<0.1; **p<0.05; ***p<0.01

Aspiration, in the philosophical sense, is a much more ambiguous latent construct for which open-ended responses may be better suited. In the mean differences reported in Table 9 we find that hosts tend to have higher secular aspirations, and refugees have higher levels of religious aspiration, and that parents tend to have higher levels of both kinds of aspiration for male children. The regression results in Table 11 continue to show that a parent tends to have more clearly articulated aspirations for a male child. Better educated parents are more likely to express secular aspirations, and parent's aspirations tend to be lower for older children. The host-refugee differences are no longer significantly different from each other.

Table 11: Aspiration and household characteristics

	Dependent variable:					
	Secular Aspirations		Religious Aspirations		Religious Education	
	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)
Refugee status	−0.018	−0.017	−0.011	−0.015	0.015	−0.006
	(0.022)	(0.012)	(0.015)	(0.010)	(0.016)	(0.010)
Eldest child's sex	−0.015	−0.014*	−0.025***	−0.034***	−0.017*	−0.027***
	(0.014)	(0.007)	(0.009)	(0.006)	(0.010)	(0.006)
Eldest child's age	−0.002	−0.002**	−0.002*	−0.002***	−0.003**	−0.003***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female household head	0.013	−0.011	−0.020	0.0002	−0.019	−0.001
	(0.019)	(0.010)	(0.013)	(0.009)	(0.014)	(0.008)
Number of children	0.010*	0.003	0.004	0.005**	0.002	0.004*
	(0.006)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)
Parent's years of education	0.003	0.003**	0.0001	−0.001	−0.001	−0.001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Parent's religious education	−0.007	0.021	0.033	0.011	0.012	0.028*
	(0.036)	(0.019)	(0.024)	(0.016)	(0.026)	(0.015)
Asset Index	−0.002	−0.001	−0.006	−0.002	0.003	0.002
	(0.006)	(0.003)	(0.004)	(0.003)	(0.004)	(0.002)
Household Income	0.003	−0.0001	−0.001	−0.001	−0.006	−0.002
	(0.006)	(0.002)	(0.004)	(0.001)	(0.004)	(0.001)
Trauma Event Score	−0.001	0.0001	0.002	0.002	−0.001	0.0002
	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Constant	0.067**	0.071***	0.059***	0.058***	0.085***	0.076***
	(0.033)	(0.017)	(0.023)	(0.014)	(0.024)	(0.014)
Observations	311	842	311	842	311	842
R^2	0.026	0.026	0.070	0.048	0.056	0.046
Adjusted R^2	−0.007	0.014	0.039	0.037	0.025	0.034
Residual Std. Error	0.119	0.106	0.080	0.089	0.085	0.087
F Statistic	0.791	2.202**	2.251**	4.224***	1.785*	3.992***
Note:						*p<0.1; **p<0.05; ***p<0.01

Measures of navigational capacity, which are derived from the open-ended responses to two questions "What have you done to achieve these goals?" and "What do you plan to do in the future to achieve these goals?" show interesting results. We have coded the responses to these questions into measures of "capacity" – which captures the ability to articulate a clear navigational capacity, and "budget" which codes responses related to availability of funds to achieve a particular ambition or aspiration. The mean differences in Table 9 show that refugees have more clearly articulated navigational capacity (particularly in the enhanced sample), but less budget. The regression results in Table 12 provide more insight into these mean differences – the refugees continue to show much higher navigational capacity, and the host effect on budget survives the introduction of additional controls. We think that the refugee effect on capacity is a selection and learning effect, given the extraordinary efforts that refugee families in Cox's bazaar have taken to survive the violence in Myanmar and find their way to the camp. The regressions also show that more educated parents demonstrate both higher capacity and higher budget. And, not surprisingly, household income is positively correlated with higher budget. Finally, the regressions in Table 13 show clear

gender differences in ambitions for marriage and migration. Parents are much more likely to talk about marriage as a goal for their female children, and migration as a goal for their male children. Discussion about marriage tends to get tempered with the age of child.

We would also note that the coefficients in the enhanced sample regressions generally do not differ from the human-coded sample. This is encouraging because it shows that, once we control for household characteristics, the enhanced and human sample do not appear to be biased, and any differences between them are due to the differences in sample size.

Table 12: Navigational Capacity and household characteristics

	Dependent variable:			
	Capacity		Budget	
	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)
Refugee status	0.110	0.146***	0.033	0.090
	(0.078)	(0.049)	(0.087)	(0.057)
Eldest child's sex	0.033	−0.010	0.047	0.023
	(0.049)	(0.031)	(0.054)	(0.036)
Eldest child's age	−0.005	−0.0003	−0.022***	−0.011**
	(0.006)	(0.004)	(0.007)	(0.005)
Female household head	0.031	−0.026	0.038	−0.058
	(0.068)	(0.042)	(0.073)	(0.048)
Number of children	0.038*	0.013	0.042*	0.003
	(0.021)	(0.013)	(0.025)	(0.015)
Parent's years of education	0.023***	0.016***	0.034***	0.016***
	(0.007)	(0.005)	(0.008)	(0.005)
Parent's religious education	−0.044	0.080	−0.084	−0.082
	(0.143)	(0.079)	(0.129)	(0.088)
Asset Index	0.008	−0.007	0.010	0.006
	(0.021)	(0.013)	(0.023)	(0.016)
Household Income	0.026	0.007	0.048**	0.042***
	(0.020)	(0.006)	(0.023)	(0.014)
Trauma Event Score	0.005	−0.009	−0.005	−0.007
	(0.010)	(0.006)	(0.012)	(0.007)
Constant	0.126	0.345***	0.218	0.318***
	(0.119)	(0.072)	(0.144)	(0.087)
Observations	265	681	196	493
R^2	0.077	0.042	0.212	0.085
Adjusted R^2	0.040	0.028	0.170	0.066
Residual Std. Error	0.389	0.398	0.370	0.396
F Statistic	2.107**	2.958***	4.991***	4.489***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13: Marriage, Migration, Camp Regulations and household characteristics

	Dependent variable:					
	Marriage		Migration		Camp Regulations	
	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)	(*Human*)	(*Enh.*)
Refugee status	0.046^{*}	0.019	0.013	0.009	0.040^{***}	0.022^{***}
	(0.024)	(0.015)	(0.012)	(0.008)	(0.012)	(0.005)
Eldest child's sex	0.082^{***}	0.110^{***}	-0.023^{***}	-0.019^{***}	−0.005	−0.004
	(0.015)	(0.010)	(0.007)	(0.005)	(0.007)	(0.003)
Eldest child's age	0.007^{***}	0.005^{***}	0.001	0.001	0.0002	−0.0001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)
Female household head	0.018	0.016	-0.018^{*}	-0.012^{*}	−0.009	−0.001
	(0.021)	(0.013)	(0.010)	(0.007)	(0.010)	(0.005)
Number of children	−0.005	0.002	0.001	−0.002	0.002	0.002
	(0.006)	(0.004)	(0.003)	(0.002)	(0.003)	(0.001)
Parent's years of education	−0.003	−0.001	0.001	−0.001	−0.002	−0.001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Parent's religious education	−0.041	-0.059^{**}	0.019	0.001	−0.026	−0.013
	(0.039)	(0.024)	(0.019)	(0.012)	(0.020)	(0.009)
Asset Index	−0.006	−0.003	−0.001	−0.001	0.003	0.001
	(0.006)	(0.004)	(0.003)	(0.002)	(0.003)	(0.001)
Household Income	0.009	−0.0002	−0.001	−0.001	0.003	0.0002
	(0.006)	(0.002)	(0.003)	(0.001)	(0.003)	(0.001)
Trauma Event Score	0.0001	0.0003	−0.0001	−0.001	0.0002	0.001
	(0.003)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Constant	−0.034	−0.035	0.005	0.030^{***}	−0.004	−0.003
	(0.036)	(0.022)	(0.018)	(0.011)	(0.018)	(0.008)
Observations	311	842	311	842	311	842
R^2	0.160	0.174	0.073	0.034	0.089	0.062
Adjusted R^2	0.132	0.164	0.043	0.023	0.059	0.051
Residual Std. Error	0.129	0.136	0.063	0.069	0.065	0.048
F Statistic	5.713^{***}	17.554^{***}	2.379^{**}	2.942^{***}	2.944^{***}	5.536^{***}
Note:						*p<0.1; **p<0.05; ***p<0.01

7 Conclusion

We have argued that the narrative responses obtained from qualitative, open-ended questions are potentially important to explore concepts that are of central interest to economists and policy makers. They are underused mainly because they are considered "anecdotal" since they are hard to analyze for larger samples. Qualitative responses tend to be collected by the other social sciences for small samples because of the intense effort required to code and analyze them. To address this, we have begun the development of a method that employs Natural Language Processing to expand a subset of human-coded responses to a larger, more representative, "enhanced" sample and thus bridge the gap between small-sample qualitative work and large-sample quantitative work. This paper is an initial step in the development of this method, which we hope to refine in subsequent work.

We apply this method to the widely studied question of aspirations, bringing in ideas from

philosophy and anthropology to distinguish between "aspirations" (a process of reversing a "core value" that results in a "change in the self") and "ambition" (a specific goal that can be "fully grasped in advance of achieving it"), and to include the concept of the "capacity to aspire" which is a cultural and cognitive resource that allows people to navigate their way to a better future. We collect narrative data from open-ended questions asked to a representative sample of Rohingya refugees and their Bangladeshi hosts in Cox's Bazaar in Bangladesh. We find that there are interesting and policy-relevant differences between general and educational ambitions, and secular and religious aspirations. We also find that while poorer and less educated parents (more prevalent among the Rohingya refugees) say that budget constraints limit their ability to help their children achieve their hopes and aspirations for them, refugees are more able to express clearly articulated navigational capacity than hosts, which we think may be due to the selection and learning effects associated with surviving violence and finding their way to the refugee camp.

In the application we present here, we extend 395 human annotations to a sample of 1,040, so more than double the sample size. We show that this increases the precision of estimates and therefore allows the identification of patterns that are not observable in the smaller sample. However, we expect our method to be even more useful in larger samples and we are currently working on applying our approach to a larger set of interviews from a later round of the Cox's Bazaar survey. This larger sample will also allow us to explore questions such as well-being that may also be amenable to narrative analysis.

In our future work, we also plan to explore possible data augmentation methods such as back translation to increase the effective sample size of our training sets, thus potentially allowing the use of richer supervised classification models. We are working on a Python package that implements our method and hope to make it publicly available in the near future.

References

Ahrens, M., Ashwin, J., Calliess, J.-P., and Nguyen, V. (2021). Bayesian topic regression for causal inference. *arXiv preprint arXiv:2109.05317*.

Alexander, J. T., Andersen, R., Cookson Jr, P. W., Edin, K., Fisher, J., Grusky, D. B., Mattingly, M., and Varner, C. (2017). A qualitative census of rural and urban poverty. *The Annals of the American Academy of Political and Social Science*, 672(1):143–161.

Apel, M. and Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series No. 92*.

Appadurai, A., Vijayendra, R., and Michael, W. (2004). The capacity to aspire: Culture and the terms of recognition. *Culture and Public Action, ed. Vijayendra Rao and Michael Walton, Stanford, California: Stanford University Press*, pages 59–84.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Berezkin, Y. E. (2015). Folklore and mythology catalogue: its lay-out and potential for research. *The Retrospective Methods Network*, (S10):58–70.

Blei, D. M. and McAuliffe, J. D. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128.

Callard, A. (2018). *Aspiration: The agency of becoming*. Oxford University Press.

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., and Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–20.

Correa, R., Garud, K., Londono, J. M., Mislang, N., et al. (2017). Constructing a dictionary for financial stability. *IFDP notes. Board of Governors of the Federal Reserve System, Washington, DC*.

Crowston, K., Liu, X., and Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. *proceedings of the American Society for Information Science and Technology*, 47(1):1–2.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

Espeland, W. N. and Stevens, M. L. (1998). Commensuration as a social process. *Annual review of sociology*, 24(1):313–343.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Ferguson-Cradler, G. (2021). Narrative and computational text analysis in business and economic history. *Scandinavian Economic History Review*, pages 1–25.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115–132.

Fruttero, A., Muller, N., and Calvo-Gonzalez, O. (2021). The power and roots of aspirations. Technical report.

Genicot, G. and Ray, D. (2017). Aspirations and inequality. *Econometrica*, 85(2):489–519.

Genicot, G. and Ray, D. (2020). Aspirations and economic behavior. *Annual Review of Economics*, 12.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Jayachandran, S., Biradavolu, M., and Cooper, J. (2021). Using machine learning and qualitative interviews to design a five-question women's agency index. Technical report, Northwestern Global Poverty Research Lab Working Paper.

Karamshuk, D., Shaw, F., Brownlie, J., and Sastry, N. (2017). Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, 1:33–43.

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520.

Liew, J. S. Y., McCracken, N., Zhou, S., and Crowston, K. (2014). Optimizing features in active machine learning for complex qualitative content analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 44–48.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Mann, K. and Püttmann, L. (2018). Benign effects of automation: New evidence from patent texts. *Available at SSRN 2959584*.

Michalopoulos, S. and Xue, M. M. (2021). Folklore. *The Quarterly Journal of Economics*, 136(4):1993–2046.

Nimark, K. P. and Pitschner, S. (2019). News media and delegated information choice. *Journal of Economic Theory*, 181:160–196.

Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127:104119.

Parthasarathy, R., Rao, V., and Palaniswamy, N. (2019). Deliberative democracy in an unequal world: A text-as-data study of south india's village assemblies. *American Political Science Review*, 113(3):623–640.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rao, V. (2022). Can economics become more reflexive? exploring the potential of mixed-methods. *World Bank Group Policy Research Working Paper*, (9918).

Ray, D. (2006). Aspirations, poverty, and economic change. *Understanding poverty*, 1:409–421.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Saiz, L., Ashwin, J., Kalamara, E., et al. (2021). Nowcasting euro area gdp with news sentiment: a tale of two crises. *ECB Working Paper Series*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

Shiller, R. J. (2020). *Narrative economics*. Princeton University Press.

Small, M. L. (2009). How many cases do i need?' on science and the logic of case selection in field-based research. *Ethnography*, 10(1):5–38.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Veitch, V., Sridhar, D., and Blei, D. (2020). Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Wiedemann, G. (2019). Proportional classification revisited: Automatic content analysis of political manifestos using active learning. *Social Science Computer Review*, 37(2):135–159.

World Bank (2019). Cox's bazaar - baseline survey (april 2019-august 2019): Baseline information document. Technical report, Poverty Global Practice, World Bank.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Yordanova, K. Y., Demiray, B., Mehl, M. R., and Martin, M. (2019). Automatic detection of everyday social behaviours and environments from verbatim transcripts of daily conversations. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, pages 1–10. IEEE.

A Coding Tree

Table 14: Definitions and Examples from transcripts of Aspiration

Code	**Subcode**	**Definition**	**Examples from transcripts**
Aspiration	Religious	Religiously motivated aspirations for children.	Expressions of parental desires for their children that were coded for religious aspirations: • Ability to read Quran • Maintain Islamic covering • Prays regularly or Prays 5 times • Works in Islamic banks • Become a maulvi / alem / alemdar / elamdar / mawlana [i.e equivalent to an Islamic Scholar] • Become hafiz / hafez [i.e memorize Quran] or wants to send to hafez khana [i.e send to schooling that primarily focuses on helping children memorize Quran] • Send to noorani madrassa / school [i.e schooling for religious education equivalent to primary level] • Wants to send to madrassa [i.e attend schooling which follows religious curriculum] • Wants the child to learn/study Arabic
Aspiration	Secular	Expressions of parental aspirations in terms of positive character traits, which can be intangible, or desire for unspecified positive things to happen to the child (e.g., hoping for a good life partner for the child or hoping the child to attain decent standard of living).	Expressions of parental desires for their children that were coded for secular aspirations: • Take care of wife and children and old parents by doing jobs • Earn enough money to live a beautiful life • Be healthy and have a respectable job • If people recognize him [give him recognition] • Earn well and build a house • The more prosperous my child gets, the happier I will be. • Make him a doctor for the good of the nation

Table 15: Definitions and Examples from transcripts of Ambition

Code	Subcode	Definition	Examples from transcripts
Ambition	No Ambition	Expressions of helplessness in context of ambitions or implied unwillingness to, or lack of dream/plan.	• There is nothing to do except sitting quietly. • I have no hope • There is no plan because I don't understand • No hope for girls, they will get married
Ambition	Vague - Job	Only coded when job type was unspecified in the context of ambition	When parents want child to • Make money • get a good job • get a job • Work hard • earn well
Ambition	Job Secular	Coded when specific job, occupation or work type was highlighted.	Doctor, Government job, NGO job, Teacher in non-religious school
Ambition	Vocational Training	Any vocational training in the context of ambition is mentioned.	Tailoring/Handicrafts
Ambition	Job Secular	Coded when specific job, occupation or work type was highlighted.	Doctor, Government job, NGO job, Teacher in non-religious school
Ambition	Entrepreneur	Coded when non-wage enterprise job is mentioned. Applies regardless of whether business type is specified.	Shopkeeper, business, own farm
Ambition	Education Low	Coded when dreams for the child's education are lower or equivalent to higher secondary (for non-religious education) or noorani madrassa (for religious education). The code is not used if parent indicates the current status of the child, e.g., "my child is studying at class 10". For the code to apply, it has to be a future ambition Also, code is not used if the education not specific, e.g., "I want to teach my child Arabic."	• I hope to educate him up to tenth grade. • I had hoped to educate her up to SSC but now I cannot educate her due to the lack of money.
Ambition	Education Neutral	Coded when education is mentioned in vague terms. Also coded when 'madrassa' is referred as a religious education ambition.	• I hope to get the boy educated till the end. • If he wants to study, then I will educate him as long as he wants to.
Ambition	Education High	Coded when dreams for the child's education are above higher secondary (for non-religious education) or for high religious education.	• I want my child to study engineering. • I want my child to be a maulvi.
Ambition	Education Religious	Coded along with all Aspiration:Religious aspiration codes aside from when hafezi is mentioned. However, code also if "sending to hafez khana" is a future dream	My child will become a: • Maulvi / Alem / Alemdar / Elamdar / Mawlana My child will go to: • noorani madrassa/school • madrassa • Hafez khana • learn arabic
Ambition	Marriage	Coded any time marriage is mentioned in the context of ambition	• will get her married
Ambition	Migration	Any time ambition is related to leaving current place of residence for work, studying or resettling.	• Go abroad • Go back to Burma

Table 16: Definitions and Examples from transcripts of Navigational Capacity

Code	Subcode	Definition	Examples from transcripts
Navigational Capacity	Vague/Non Specific	When parent mentioned unspecific or unclear attempts/measures to help achieve dreams for child.	• trying hard • will do as much as I can • will do my best • let's see what happens
Navigational Capacity	Reliance on God	When either the parent fully/partially relies on God to fulfill future dream for children or is fully/partially reliant on God at present.	• even if there is hope, it depends on God willing • god is running our lives somehow
Navigational Capacity	Ability High	Coded when the parent demonstrates having gone the extra mile ensure a better future for the child. This needs to be coded inferentially, as no specific sequence of repeating words/phrases can be strictly identified to classify instances of high ability.	• I am somehow managing my children's education by borrowing money from my brothers. • We try to cover our expenditures by selling some of the items from the monthly aid that we get. [Double coded with Budget Low]
Navigational Capacity	Ability Low	Coded when the parent specified having no resources to help the child.	• What can we do from here? We are having to stay how we are.
Navigational Capacity	Budget High	Coded when the parent expresses having money, including an ability to save or spend money.	• I am educating her anyway I can. By helping financially, with hard work, appointing a private tutor and financing their education.
Navigational Capacity	Budget Low	Coded when the parent expresses not having money.	• Hoping to teach her as per the ability Allah grants me. However, if there is money involved, I cannot educate her.
Navigational Capacity	Awareness Information High	Coded when the parent displays awareness or information. Inferentially coded.	• I talk to my husband, so that he doesn't obstruct the children's education in any way. There is nothing to do here without education. If they do not study, their future will be dark. To brighten their future, they have to be educated in any way. We had places and properties when we were in Myanmar. But now, we don't have anything here, except to study. That's why I am trying to educate my children. [Double coded with High Ability]
Navigational Capacity	Awareness Information Low	Not knowing what to do, clueslessness.	• Question: What kind of doctor would you be happy with? Answer: He could be a popular doctor.

B Measurement errors

Table 17: Measurement error variances (bootstrapped with replacement)

Code	N_h	N_m	$\hat{\sigma}^2_h$	$\hat{\sigma}^2_m$	$\hat{\sigma}^2_\epsilon$	$\hat{se}_h$	$\hat{se}_m$	$\hat{se}_{enh}$
Secular Aspiration	395	645	0.015	0.009	0.009	0.006	0.005	0.004
Religious Aspiration	395	645	0.007	0.007	0.005	0.004	0.004	0.003
Religious Education	395	645	0.007	0.005	0.006	0.004	0.004	0.003
General Ambition	387	645	0.108	0.099	0.078	0.017	0.017	0.012
Education Ambition	355	637	0.115	0.099	0.108	0.018	0.018	0.013
Capacity	340	640	0.153	0.169	0.164	0.021	0.023	0.017
Budget	257	564	0.162	0.168	0.085	0.025	0.021	0.017
Marriage	395	645	0.021	0.024	0.004	0.007	0.007	0.005
Migration	395	645	0.007	0.004	0.004	0.004	0.004	0.003
Camp Regulations	395	645	0.004	0.001	0.004	0.003	0.003	0.002

Table 18 shows the results for regressions of prediction errors for the 3-fold cross validated model using English tf-idf vectors on the full annotated sample. A positive coefficient on, for example, refugees here means that refugees are more likely to have a positive prediction error compared to hosts. This would suggest that the machine annotations for refugees may be too low (i.e. too negative). The F statistics from this table are reproduced in the main body of the paper in Table 6.

C Robustness

Table 20 shows the correlation of enhanced sample variances using a range of different pre-trained word embedding vectors as the model inputs, as well as tf-idf vectors on the original Bengali transcripts. We see that in all cases the correlation with the baseline is over 0.7, with the exception of tf-idf vectors in the original Bengali which have particularly bad predictive performance. Table 21 shows the variance of the out-of-sample prediction errors for each embedding, compared to the baseline mode. This variance is crucial as it both gives us an indication of model fit and is used in adjustments for the measurement error variance. The embeddings are described in more detail in Table 19 below. All embeddings were computed using the sentence-transformers Python package,[17] so are converted to sentence level scores using the method proposed by Reimers and Gurevych (2019).

[17] https://github.com/UKPLab/sentence-transformers

Table 18: Regressing prediction errors of qualitative codes on household characteristics

	Prediction error for:									
	Secular Aspiration	Religious Aspiration	Religious Education	General Ambition	Education Ambition	Capacity	Budget	Marriage	Migration	Camp Regulation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Refugee status	0.019	−0.011	−0.0001	0.052	0.066*	−0.026	0.026	0.004	0.004	−0.010
	(0.013)	(0.011)	(0.009)	(0.042)	(0.039)	(0.063)	(0.047)	(0.010)	(0.007)	(0.008)
Eldest child's sex	0.0004	−0.004	−0.007	−0.032	−0.030	−0.019	0.0003	−0.016***	−0.003	−0.002
	(0.008)	(0.007)	(0.005)	(0.025)	(0.024)	(0.040)	(0.029)	(0.006)	(0.004)	(0.005)
Eldest child's age	−0.0004	−0.001	−0.001	0.001	0.006	−0.005	−0.005	−0.001	−0.0004	−0.001
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.005)	(0.004)	(0.001)	(0.001)	(0.001)
Female household head	0.003	−0.003	0.001	0.022	−0.060*	0.081	0.044	−0.003	0.004	0.003
	(0.011)	(0.010)	(0.008)	(0.036)	(0.034)	(0.057)	(0.038)	(0.008)	(0.006)	(0.007)
Number of children	0.005	−0.001	0.004	0.004	−0.003	0.022	0.019	0.001	0.001	0.006***
	(0.003)	(0.003)	(0.002)	(0.011)	(0.011)	(0.017)	(0.013)	(0.003)	(0.002)	(0.002)
Parent's years of education	0.0004	0.0004	−0.001	0.002	0.008**	0.0004	0.007	−0.0005	−0.001	−0.0003
	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.006)	(0.005)	(0.001)	(0.001)	(0.001)
Parent's religious education	−0.010	0.034*	−0.012	−0.080	0.116*	0.047	−0.020	−0.029*	−0.005	0.001
	(0.021)	(0.018)	(0.014)	(0.068)	(0.065)	(0.112)	(0.070)	(0.016)	(0.011)	(0.013)
Asset Index	−0.001	−0.004	0.002	0.022**	0.015	0.009	0.001	−0.0003	0.003	−0.003
	(0.003)	(0.003)	(0.002)	(0.011)	(0.010)	(0.017)	(0.013)	(0.003)	(0.002)	(0.002)
Household Income	0.002	−0.004	−0.002	0.001	−0.011	0.002	−0.009	0.003	0.0001	0.005**
	(0.003)	(0.003)	(0.002)	(0.011)	(0.011)	(0.017)	(0.012)	(0.003)	(0.002)	(0.002)
Trauma Event Score	0.0001	0.0005	0.001	−0.002	−0.005	0.002	−0.009	0.001	0.001	0.002
	(0.002)	(0.001)	(0.001)	(0.005)	(0.005)	(0.009)	(0.006)	(0.001)	(0.001)	(0.001)
Constant	−0.017	0.027	0.007	−0.003	0.016	0.0002	0.061	−0.006	−0.004	−0.024**
	(0.019)	(0.016)	(0.013)	(0.061)	(0.058)	(0.098)	(0.076)	(0.014)	(0.010)	(0.012)
Observations	311	311	311	298	274	242	180	311	311	311
R^2	0.036	0.035	0.030	0.035	0.064	0.019	0.046	0.050	0.022	0.061
Adjusted R^2	0.004	0.002	−0.002	0.001	0.029	−0.024	−0.010	0.018	−0.010	0.029
Residual Std. Error	0.069	0.059	0.047	0.214	0.196	0.303	0.190	0.052	0.036	0.042
F Statistic	1.123	1.074	0.933	1.033	1.804*	0.435	0.819	1.571	0.681	1.933**

Note: *p<0.1; **p<0.05; ***p<0.01

Table 19: Embedding model description

Embedding	Description
LaBSE	Language-Agnostic BERT Sentence Embedding: a multilingual BERT embedding model, that produces language-agnostic cross-lingual sentence embeddings for 109 languages (Feng et al., 2020)
MPNet (base)	Masked and Permuted Pre-training for Language Understanding: a pre-training method for language understanding tasks (Song et al., 2020).
MPNet (paraphrase)	The MPNet model re-trained on the Paraphrase and Semantic Similarity in Twitter dataset (Xu et al., 2015).
MPNet (multilingual)	The MNet model converted to a multilingual model following Reimers and Gurevych (2020).
GloVe	Global Vectors for Word Representation: a classic unsupervised learning algorithm for obtaining vector representations for words (Pennington et al., 2014)
DistilBERT	A smaller, faster, general purpose language representation model, based on BERT (Sanh et al., 2019)

Table 20: Correlation of enhanced sample with baseline case

Text representation	Language	Secular Aspiration	Religious Aspiration	Religious Education	General Ambition	Education Ambition	Capacity	Budget	Marriage	Migration	Camp Regulation
LaBSE	Bengali	0.796	0.825	0.739	0.825	0.803	0.740	0.864	0.953	0.932	0.731
LaBSE	English	0.775	0.900	0.779	0.823	0.821	0.732	0.900	0.973	0.927	0.769
MPNet (base)	English	0.805	0.872	0.776	0.789	0.771	0.684	0.825	0.945	0.864	0.873
MPNet (paraphrase)	English	0.790	0.884	0.753	0.810	0.780	0.755	0.845	0.967	0.907	0.793
MPNet (multilingual)	English	0.819	0.907	0.779	0.803	0.825	0.746	0.873	0.963	0.932	0.830
GloVe	English	0.804	0.837	0.702	0.715	0.676	0.704	0.711	0.913	0.882	0.769
DistilBERT	English	0.847	0.844	0.748	0.801	0.809	0.741	0.873	0.979	0.916	0.809
tf-idf	Bengali	0.523	0.629	0.694	0.732	0.704	0.726	0.809	0.494	0.649	0.688

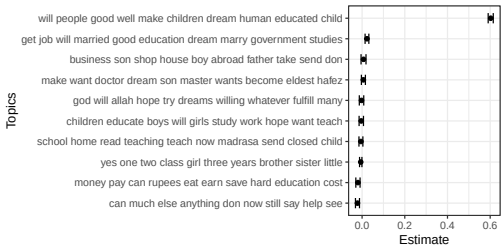
Table 21: Prediction error variance $\hat{\sigma}_\epsilon$ compared to the baseline

Text representation	Language	Secular Aspiration	Religious Aspiration	Religious Education	General Ambition	Education Ambition	Capacity	Budget	Marriage	Migration	Camp Regulation
Baseline		0.011	0.004	0.007	0.072	0.106	0.155	0.101	0.003	0.003	0.004
LaBSE	Bengali	0.008	0.004	0.008	0.070	0.110	0.169	0.063	0.011	0.002	0.004
LaBSE	English	0.009	0.005	0.006	0.072	0.075	0.196	0.038	0.003	0.003	0.004
MPNet (base)	English	0.008	0.007	0.007	0.052	0.078	0.177	0.112	0.007	0.003	0.005
MPNet (paraphrase)	English	0.009	0.006	0.006	0.060	0.076	0.188	0.089	0.005	0.002	0.006
MPNet (multilingual)	English	0.009	0.004	0.006	0.059	0.090	0.195	0.079	0.005	0.002	0.003
GloVe	English	0.010	0.005	0.008	0.082	0.100	0.165	0.155	0.008	0.004	0.005
DistilBERT	English	0.011	0.006	0.006	0.063	0.074	0.178	0.071	0.004	0.002	0.004
tf-idf	Bengali	0.028	0.006	0.009	0.103	0.106	0.177	0.095	0.039	0.009	0.005

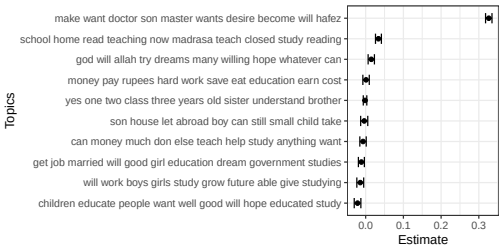
D Supervised LDA models

Figure 6: Supervised LDA on model predictions

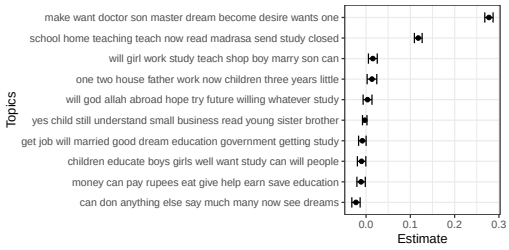
(a) Secular Aspiration



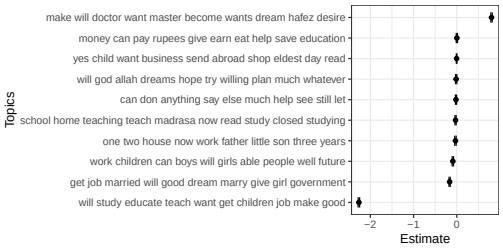
(b) Religious Aspiration



(c) Religious Education

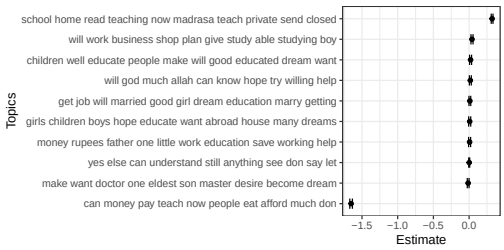


(d) Education Ambition

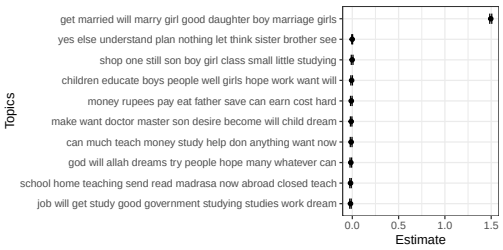


(e) Budget

(f) Marriage



(g) Migration



(h) Camp Regulations