Policy Research Working Paper 10046

A Method to Scale-Up Interpretative Qualitative Analysis, with an Application to Aspirations in Cox's Bazaar, Bangladesh

Julian Ashwin
Vijayendra Rao
Monica Biradavolu
Aditya Chhabra
Arshia Haque
Afsana Khan
Nandini Krishnan

WORLD BANK GROUP
Development Economics
Development Research Group
May 2022

Policy Research Working Paper 10046

Abstract

The qualitative analysis of open-ended interviews has vast potential in economics but has found limited use. This is partly because the interpretative, nuanced human reading of text and coding that it requires is labor intensive and very time consuming. This paper presents a method to simplify and shorten the coding process by extending a small set of interpretative human-codes to a larger, representative, sample using natural language processing and thus analyze qualitative data at scale. It applies it to analyze 2,200 open-ended interviews on parent's aspirations for children with Rohingya refugees and their Bangladeshi hosts. It shows that studying aspirations with open-ended interviews extends the economics focus on material goals to ideas from philosophy and anthropology that emphasize aspirations for moral and religious values, and the navigational capacity to achieve these aspirations. The paper shows how to assess the robustness and reliability of this approach and finds that extending the sample of interviews, rather than the human-coded training set, is likely to be optimal.

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.

A Method to Scale-Up Interpretative Qualitative Analysis, with an Application to Aspirations in Cox's Bazaar, Bangladesh*

Julian Ashwin, Vijayendra Rao†, Monica Biradavolu, Aditya Chhabra, Arshia Haque, Afsana Khan, Nandini Krishnan

Originally published in the Policy Research Working Paper Series on *May 2022*. This version is updated on *February 2023*.
To obtain the originally published version, please email prwp@worldbank.org.

*This paper supersedes our previous paper "Qualitative Analysis at Scale: An Application to Aspirations in Cox's Bazaar, Bangladesh." We have developed an open-source Python package to use the methods developed in this paper which is available here: https://github.com/worldbank/iQual. We would like to thank participants at the CSAE Lunchtime Seminar, Methods and Measurement Conference 2021, the World Bank's "Half-Baked" seminar, the October 2022 meeting of the CIFAR Boundaries, Membership and Belonging Program, the LSE Inequalities Seminar, and Ikechi Okorie for their useful comments and feedback. Peer Nagi, Eleni Kalamara and Sudarshan Aittreya 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.

†Corresponding author: Development Research Group, The World Bank, 1818 H street NW, Washington DC 20433, vrao@worldbank.org

1 Introduction

Economists almost never analyze qualitative data. We typically analyze quantitative data from structured survey questions because they are easier to administer to large representative samples of respondents, and easier to analyze using standard econometric methods. However, many questions of interest to economists may be better captured with open-ended qualitative interviews rather than structured questionnaires. These include important concepts like well-being, social norms, cultural change, vulnerability, resilience, decision-making, processes of change in interventions and experiments, and – the focus of this paper – aspirations. Structured questions work best on concepts where the possible range of responses, and follow-up questions, can be predicted in advance by the researcher. They also require that respondents have the same understanding of the latent construct underlying the question as the researcher.

For these reasons, structured quantitative 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 their range of responses cannot be anticipated in advance. When structured questions require responses with a number, or a selection from a set of choices, they can result in metrics that have the appearance of being clearly defined but hide the complexity of the "truth" (Espeland and Stevens, 1998). Latent constructs that are more subtle and nuanced are, therefore, arguably better studied with open-ended questions where the respondent is allowed the freedom to respond in an open-ended conversational style and in the manner of their choosing, and where a trained interviewer can probe an issue in a relatively unstructured manner by iteratively asking follow-up questions in a more conversational style. This process also has the advantage of eliciting information that is more "bottom-up" and driven by the respondent rather than designed ex-ante by the researcher.

Open-ended approaches to interviews have not been employed much by economists because analyzing them is hard and almost impossible to do at scale with statistically representative samples (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, reading, interpreting, and carefully coding them within the context of a theory or conceptual framework. Coding is a labor-intensive process typically done by trained social scientists,

and is an essential step in conducting nuanced analysis of qualitative data that is based on human interpretation. Interpretative qualitative analysis is consequently associated with very small sample studies. This small sample challenge that has been intrinsic to qualitative methods has resulted in a large methodological literature on qualitative and case-study methods focusing on justifying and interpreting data from interviews gathered from samples that are not designed to be statistically representative of larger populations. Their general approach 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 with statistically representative samples by combining interpretative 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 trained coders who read the transcripts, decide on a "coding-tree" and then code the transcripts using qualitative analysis 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 is then analyzed using standard regression analysis. The methods developed in our paper are not as much a major advance in Natural Language Processing (NLP) and Machine Learning methods as they are a practical contribution to the menu of tools available to economists and social scientists, and to extensively test the robustness and reliability of this approach. Our methods allow social scientists to analyze representative samples of open-ended qualitative interviews, and to do so by inductively creating a coding structure that emerges from a close, human reading of a sub-sample of interviews that are then used to predict codes on the larger sample. We see this as an organic extension of traditional, interpretative, human-coded qualitative analysis, but done at scale.

This method has several advantages over "unsupervised" NLP 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 inductively 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-defined

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" NLP 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 believe the method has wide applicability for a variety of questions of interest to economists. In this paper we apply it to study parents' aspirations for their children by analyzing data from open-ended interviews conducted on a sample of approximately 2,200 Rohingya refuges and their Bangladeshi hosts in Cox's Bazaar, Bangladesh.[1]

Aspirations are an interesting subject to apply this method, because an open-ended approach allows us to study dimensions of aspirations that are difficult to capture in structured questionnaires. The literature on aspirations in development economics (Fruttero et al., 2021) focuses on what the philosopher Agnes Callard (2018) has called "ambition" - specific goals that parents may have for their children such as a level of education, or a profession. Open-ended interviews allow us to expand this to explore its moral and spiritual dimensions - what Callard calls "aspiration" to distinguish it from "ambition" - such as being a "good person", or being religiously inclined. They also allow us to study what the anthropologist Arjun Appadurai (2004) has called the "capacity to aspire" or the capability to navigate your way to achieving a given goal. This paper applies the method we develop to differentiate between, and analyze the correlates of, ambition, navigational capacity, and aspirations (in Callard's sense) among Rohingya refugees and their Bangladeshi hosts using open-ended interviews. It demonstrates that they are independent concepts that have distinctly different determinants which suggest different policy responses.

The paper proceeds next by providing a brief overview of the literature on narrative analysis in

[1] Using basic human-coding to create a training set has been used by a few others to analyze large corpora of secondary text data (e.g. Bonikowski and DiMaggio (2022)). Our contribution in this paper is develop a method to use iterative, inductive and relatively nuanced human-coding, typical of reflexive qualitative research, to train a representative sample of primary qualitative data collected by the authors. While the sample is much larger than those analyzed in standard qualitative analysis, it is much smaller than the tens of thousands of text documents usually analyzed by existing NLP methods. This presents us with some small-sample challenges for NLP that we have attempted to resolve (Bonikowski and Nelson, 2022). Coding packages such as Atlas-TI can be used to compare the "thematic proximity" between themes identified by qualitative analysis Armborst (2017). While this approach uses the annotations provided by qualitative analysis, it does not expand the size of the annotated sample as we are proposing.

economics and aspirations in development economics, as well as placing this paper in the context of the natural language processing literature. 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. In Section 5 we then move to the NLP methodology for extending the human coded sample, describing how we cross-validate over text representations and classifiers. We also include a discussion of the role of machine translation in our analysis. Section 6 then discusses a range of tests that assess the value of the enhanced sample we create: testing for bias, efficiency and interpretability. Section 7 then sets out and discusses results for both the human and enhanced samples, illustrating the added value of the sample enhancement. Section 8 then describes a series of experiments in which we assess how the number oh human and machine annotated documents affect results. We find that for researchers on a limited budget, partially machine annotating their sample is likely to be optimal. Finally, Section 9 concludes and makes suggestions for further work. We have developed an open-source Python package called iQual (for Interpretative Qualitative Analysis) that will facilitate the use of the method[2]

2 Narrative Analysis and Aspirations

Narrative Analysis in Economics. The difficulties with using qualitative methods 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).[3]

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. The use of

[2]https://github.com/worldbank/iQual.

[3]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.

unsupervised statistical models to reduce the dimensionality of text documents into a set of interpretable variables that are used in further analysis; and the use of dictionary methods to extract a signal of interest from documents. An 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, and a wide range of general purpose and context-specific lexicons are available. An early example of this is Tetlock (2007) who uses a psychosocial dictionary to quantitatively measure interactions between media sentiment and the stock market. 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.[4] 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. 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.[5] Michalopoulos and Xue (2021) use an archive of manually coded motifs in folklore intro-

[4] In fact, Ashwin 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.

[5] There are also recent examples of manually annotating large samples, such as Andre et al. (2021) who use open-ended

duced 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.

There is also a related literature outside economics on training supervised models on human 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.[6] Yordanova et al. (2019) provide a good summary of the literature that focuses on predictive performance. Much of this literature aims to show that a particular modelling approach yields superior predictive performance in these tasks, but that is not our focus in this paper. To this end, we cross-validate over a wide range of both text representations and classifiers - allowing the data to determine which modelling approach is optimal in a given context. An application of this sort of approach in an economic context is (Mann and Püttmann, 2018) who use a supervised NLP model classify whether patents are related to automation. However, to the best of our knowledge, our paper represents the first attempt to demonstrate that extending samples in this way can add value in a context of open-ended interviews dealing with nuanced and complex topics that matter to social scientists.

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

survey responses to measure narratives about the macroeconomy, but rely on research assistants to annotate their entire sample.

[6]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.

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.

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) that analyzes how 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 Arjun Appadurai (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 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, 2018, 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 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, 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 government.

3 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. Our survey has three rounds: a baseline survey and then two further rounds of open-ended interviews. The baseline survey of 5,020 randomly selected households from the Cox's Bazaar population, split evenly between Rohingya and their Bangladeshi hosts, was conducted between April and August 2019 (World Bank, 2019). It consisted of two modules:

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.

The qualitative, open-ended, questions were conducted in two subsequent survey rounds in October to December 2020 and May and July 2021. We will refer to these three waves as the Round 1, Round

2 and Round 3, where Round 1 included the baseline quantitative survey, and Round 2 and Round 3 feature the open-ended interviews.

For the qualitative interviews, we attempted to obtain information from a random sample of 25% of the full sample (i.e. 1,255 households) in Round 2 and 50% of the baseline sample (i.e. 2,500 households) in Round 3. 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 in Round 2, and 2,038 interviews in Round 3. Of the 3038 interviews conducted we restrict ourselves in this analysis to households that whose eldest child lived with them and was still of school-going age. This allows for a meaningful interview on the parent's aspirations for the child, and to link the child being referred to in the open-ended interview to their individual characteristics in the baseline data. With this we lose about 901 interviews leaving us with 2177 for the analysis. Round 2 interviews on aspirations lasted around 15 minutes on average. However, to be consistent across the two rounds we Round 3 interviews were longer as they covered two additional domains, although the questions on aspiration were the same as in Round 2.[7] Both sets of open-ended interviews in the two rounds were conducted over the phone.

Interviews began with a short quantitative, 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. After extensive pre-testing and piloting, the final qualitative interview protocol that followed at the end of the short education module consisted of the following two questions:

1. Can you tell me about the hopes and dreams you have for your children?

2. What have you done to help them achieve these goals?

Round 2 qualitative data was collected by five interviewers, supervised jointly by a local survey firm and a subset of the authors of this paper. Interviewers for hosts were required to be verbally proficient in the local Chittagonian dialect, with those interviewing the Refugees were also required to be familiar with the Rohingya dialect. Interviewers who had participated in Round 2 were hired again

[7] The additional modules on well-being and inter-group relations extending the total interview duration for as long as up to 40 minutes. We leave an analysis of these additional modules for future work.

for Round 3 supplemented by additional interviewers; Round 3 data was collected by a total of 12 interviewers (5 males and 7 females). Several days of training including practice mock sessions were conducted before both rounds. The primary contents covered in the training included: i) an overview of qualitative interviews to guide interviewers on the importance of probes and the usage of respondents' own words to ask follow-up questions; ii) Qualitative Interview Ethics to reiterate principles of interviewing such as right to privacy of personal information; and iii) Probing Exercises which required each interviewer to pen down examples of "leading" versus "good" probes. Additionally, throughout the data collection in both rounds, interviewers participated in debriefing sessions, which allowed interviewers to brainstorm with the full team on appropriate interview techniques and best practices responding to any ethical challenges.

[Table 1 about here.]

A real-time dashboard demonstrating daily interview attempts, completed interviews, average interview duration and similar tracking components was developed and used to guide interviewers on pace and quality. The duration spent interviewing each of the domains was used as a quality flag. When the duration of an interview was significantly different from the average, the recordings were sent to supervisors for a thorough check. Both the supervisors and an external local language expert were each randomly assigned 5 recordings each per day to check. Their aggregated comments would later be taken to debriefs to discuss scope and specific areas of improvements.

The following pipeline was put in place to produce clean transcripts of interviews. First, 12 interviewers conducted qualitative interviews using SurveyCTO and its built-in recording features. Second, 16 transcribers prepared handwritten Bangali verbatim transcripts of the audio. Handwritten transcripts were then typed and a team listened to randomly selected audio recordings and checked for mismatches, typing and spelling errors. Third, A CATI system developed solely for uploading transcripts was used to submit the typed transcripts. The Bengali transcripts were then translated into English using the Google Translate API. A team of 12 translators appointed by the local firm were additionally used to manually translate the Bengali transcripts into English. A smaller subset was subsequently employed to correct the machine translated transcripts.

While the interviews were conducted in Bengali, we work with English translations of the transcripts. The qualitative coding described in the next Section was performed on the machine translations of the transcripts that had been manually corrected. We discuss the merits of machine vs human translation in more detail in Section **??**. Across both Round 2 and Round 3, the interviews are on average 12.6 distinct question-answer pairs long, with each answer made up of 13.7 words on average.

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.

4 Qualitative Analysis

4.1 Coding Tree

The development of the "coding tree" for the qualitative annotation exercise comprised of two distinct steps. First, co-author Vijayendra Rao [henceforth, VR] employed a concept driven or deductive approach in defining three broad categories: Aspiration, Ambition and Navigational Capacity as the primary response classification goals. For the second step, a classical inductive approach was employed by the three co-authors (Arshia Haque [henceforth, AH], Afsana Khan [AK], Monica Biradavolu [MB]) who conducted focused reading exercise on a sub sample (of 40 transcripts) in producing 21 sub-codes and their respective definitions. We ensured that this initial reading included transcripts of male and female, and host and refugee respondents to maximize the diversity in probable sub codes at a very early stage. With the annotation sample as large as 400 for each of the two rounds, the inductive approach we followed substantially improved coding efficiency in minimizing the discovery of too many new codes, and thereby the time needed to revisit previous transcripts to annotate those additions.

Using Atlas-TI, a qualitative data analysis software, a two-person team (AH and AK) coded 789 transcripts. 400 interviews 9comprising 50% host and 50% refugee) were randomly drawn from the 1,040 transcripts to be coded in Round 2. A further 400 interviews, again equally split by refugee status, were randomly drawn from the 2,040 transcripts in Round 3. Of these 800 allocated inter-

views, 11 were left uncoded due to either poor audio leading to missing data, call drops-offs, or very short responses with no plausible code applicability. Coders were asked to annotate interviews at the question-answer pair level to preserve granularity while being able to replicate the sub-division of interviews in the unannotated documents.

[Figure 1 about here.]

Figure 1 shows the coding tree. The qualitative distinctions between aspiration and ambition were adapted in this paper within the context and nature of "dreams" parents expressed for their children. For example, concrete and measurable dreams for child (e.g wishing a child would become a doctor, teacher, entrepreneur, or specific educational goals) was used as definition for ambition while intangible, value oriented goals (e.g wishing the child to live with dignity or be a good human being) was classified as aspiration. Aspirations, following Callard's definition, were divided into "Religious" and "Secular" . Ambition was divided into five major categories – education (further sub-coded into High, Low, Neutral and Religious), Salaried Employment, Marriage, Entrepreneurship, Migration, Vocational Training, and No Ambition. While ambition and aspiration came up at any point in an interview, "capacity to aspire" or Navigational capacity only appeared in response to the third question of the instrument i.e "what have parents been able to do to fulfill dreams for their children?" 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. These additional codes were for Covid Impacts, Public Assistance and Worries/Anxieties.

Descriptions and examples of these codes are displayed in Appendix A.1, but Figure 2 includes a few examples to illustrate some differences between aspirations, ambition and capacity.

[Figure 2 about here.]

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

We follow a standard approach to ensuring cross-coder agreement, with each interview being reviewed and any disagreements resolved through discussion betwen AH, AK and MB. Further details on this are given in Appendix A.2.

4.2 First look at the annotations

While we will compare the human annotated sample to our enhanced sample in detail throughout the remainder of the paper, Table 2 shows some summary statistics for the human annotations. We see that annotations are very sparse at the question-answer pair level (for example only 3.0% question-answer pairs are annotated as Religious Aspirations). However, when aggregated to the interview level there is much less sparsity (for example, 23% of interviews have at least one question-answer pair annotated as Religious Aspirations). There are also notable differences between rounds, which should not be due to differences in coding as the same coders and coding tree were used across rounds. A decrease in the question-answer pair level proportion is at least partly due to the longer interviews, but differences in the proportion of interviews with at least one positive are plausibly due to changes in circumstances/attitudes over the intervening year (which of course included a global pandemic). For example we see an increase from 14.9% of interviews in R2 mentioning Covid Impacts to 27.2% of interviews in R3.

[Table 2 about here.]

5 Methodology

In this Section, we describe the NLP modeling approach we use to scale up our sub-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 and discussion on the options for supervised models, text representations and training method we use.

In the following Section, we then explain how to assess the value and performance of this enhanced sample.

5.1 Approach

The overall goal of our approach is to use our subset of annotated interviews to provide reliable annotations for the remainder of the sample. Broadly, we do this by training a series of classifier models on our annotated set and then using this model to predict annotations for the unannotated set. More concretely, 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.

[Figure 3 about here.]

We train and predict for each of the 24 annotations separately, so the model for Religious Aspiration will be trained and make its predictions separately from the model for Secular Aspiration. Furthermore, as mentioned in Section 4, the qualitative annotations are defined at the level of question-answer pairs (QA). This allows us to represent each annotation as a binary classification problem at the QA level.

Figure 3 illustrates our overall methodology for a single annotation. On the left hand side we see a "human" sample of size N_h, in which interviews include both text w and annotations y, and a "machine" sample in which interviews include only the text. As annotations are defined at the QA (question-answer pair) level, so we represent $w^h_{i,s}$ as the sth QA in interview i in the human sample, with $y^h_{i,s}$ being the binary annotation on that QA. In other words, if the annotation Religious Aspiration, $y^h_{i,s}$ will be equal to one if that QA has been annotated as displaying religious aspirations, and will be zero otherwise.

We then train some classifier $f()$ parameterised by θ to predict $y^h_{i,s}$ based on the QA text $w^h_{i,s}$. As we will discuss below, there are many options for both the classifier we can use here, as well as how to represent the text numerically. A key point here is that the text representation must by

full unsupervised - i.e. we do not use any information about y or any further information about the interview subject when creating a numerical representation of the text. The text representation, classifier and a variety of hyperparameters are chosen using k-fold cross-validation, as we discuss in Section 5.2. Given this trained classifier we can then predict annotations a the QA level for our unannotated "machine" sample. This gives us the predicted annotations $\hat{y}^m_{i,s}$.

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: potentially picking up multiple contradictory signals within a single interview, or allowing a comparison of the frequency with which a signal appears within interviews. 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. If the documents was not in a question-answer interview form, the annotation and training could be done at the sentence or paragraph level to give similar advantages.

We then aggregate the QA level annotations to the interview level using aggregation function $g()$. The choice of this aggregation function is at least in part a substantive question that depends on the research question. For example, if we take the mean value of y across QA pairs for each interview this gives us a measure of the intensity with which this concept comes up. On the other hand, if we take the maximum value across the interview this gives us a measure of interviews in which this concept comes up at least once. We perform this aggregation for both the observed human annotations Y^h, the "in-sample" predicted human annotations $\hat{Y}^h$ and the "out-of-sample" predicted machine annotations $\hat{Y}^m$. The predicted annotations for the human sample can then by used to assess the measurement errors introduced by the model. Particularly for the quantification of measurement errors, we make extensive use of bootstrapping, but as this is conceptually separate from the core intuition of our method, we leave a discussion of this to Section 5.2. The observed human annotations and machine annotations are then combined to give an enhanced sample $\tilde{Y}$. Once we have verified that the enhancement does indeed add value, we proceed with substantive analysis.

We can then assess the value of this enhanced sample, as described in more detail in Section 6. Broadly speaking, we test whether the enhancement introduces a bias, whether it increases efficiency

(i.e. reduces standard errors) and whether it increases the interpretability of substantive analysis. This is an important step, as any interpretation of substantive results should be done with these assessments in mind. Finally, we can use our larger enhanced sample for substantive analysis, taking advantage of the larger sample size to identify effects that would not be observable with only the human annotated sample. We describe this analysis and our results around ambition, aspiration and navigational capacity in Section 7.

5.2 Modelling choices and Bootstrapping

There are many possible options for the numerical representation of the text representation w, the classifier $f()$ and the aggregation function $g()$. While the choice of aggregation function is something that we leave to the researcher's discretion, we use cross-validation to select the text representation and the classifier. As we train the classifier for each annotation independently, this allows for the fact that a different classification model of text representation may be optimal for different annotations. Appendix B.1 gives an exhaustive list of the text representations, models and hyperparameters that are selected over during cross validation.

Cross validation. As we are working with the QA pair level data we have 9,964 distinct observations in the human annotated sample, which come from the 789 annotated interviews. In our baseline case, we use the entire annotated sample as a training set and split it into three folds for cross validation.[8] We then use a combination of a grid search and the Optuna hyperparameter tuning framework Akiba et al. (2019) to choose the text representation, classifier and hyperparameters of that classifier that give the best validation set performance (as measured by the F1 score).

[Figure 4 about here.]

Text representations. In order to use the text of the QA pairs as inputs in a classifier, we need to represent them numerically. There are many possible ways to do this, but we select over several commonly used text representations. We allow the text representation to vary along three dimensions,

[8] In Section 8 we show how varying the size of the annotated sample affects performance, which allows fully out-of-sample analysis.

illustrated by the first three panels of Figure 4, which show the proportion of bootstraps in which each representation is chosen.

1. We include the answer of each QA pair only, or both the question and answer. As shown in the first panel of Figure 4, in most cases only the answer is selected. However, for some annotations like Education Neutral, both the question and answer are usually included.

2. The text representation can be based on either the English translation or a transliteration of Bengali into Latin characters. As shown in the second panel of Figure 4, in most cases the English translation performs better.

3. We select over a range of approaches to transform the text into numerical vectors. This include simple vectors based on phrase counts such as the CountVectorizer and TfidfVectorizer as well as vectors based on pre-trained word embedding models, described in detail in Appendix B.1. When using count based metrics we allow for both single and two word phrases.

By allowing selection over text representations each time we train a classifier, we account for the fact that which text representation best captures relevant text features can vary across the different annotations in our data.

Classifiers. As with numerical representations of text, there are many choices of classification model available. Our goal is not to argue for a certain model or approach to the classification task, but rather to argue for flexibility as different models will perform better in different contexts. We thus select over a range of popular classification models including logistic regression, random forests, support vector machines and neural networks, see Appendix B.1. Each of these models have a separate set of hyperparameters that are chosen through k-fold cross validation. We then compare the validation set performance of each model and choose that which performs best. Unsurprisingly, given the sparse nature of our annotations and the small training set, we find that simpler classifiers such as a random forest and logistic regression outperform larger models such as neural networks.

As we will conduct out substantive analysis at the household level, we aggregate the annotations into interview-level variables. There are of course different ways to do this, i.e. different choices of the aggregation function $g()$. For the sake of clarity, in the remainder of this paper we will use the

mean annotation value across QA pairs within an interview. This gives us a quasi-continuous measure between 0 and 1, where 0 would denote that the annotation does not appear at all in the interview and 1 would denote that every QA pair in the interview has that annotation. This mean aggregation therefore gives a measure of the intensity with which an annotation appears in an interview, while controlling for the interview's overall length.

Bootstrapping. We use two forms of bootstrapping to account for uncertainty in our predicted annotations. Firstly, when using the entire available training set we re-train the models with a different draw for the validation set split and any stochastic processes involved in training the model. Secondly, as described in more detail in Section 8, we also train models on a subset of the training data which is sampled without replacement.

By using validation set performance to select over such a wide range of text representations and classifiers, we seek to demonstrate that the specifics of the supervised model used to extend the sample are not central to our approach. In fact, we advocate being flexible over these details the optimal text representations and classifiers will differ across contexts. Our methodology is implemented in our Python package iQual.

Performance. Figure 5 shows the validation set performance, as measured by the F1 scores, across each annotation and bootstrap run.[9] As a natural benchmark, the annotation sparsity which corresponds to performance under random guesses is shown in red. In all cases our text-based models do much better than random guesses, suggesting that our enhanced sample will add value over the human annotated sample. It is worth noting however that there is considerably heterogeneity across annotations. In particular annotations that are associated with less concrete concepts, such as Awareness Information Low, No Ambition and Vague Non Specific, appear to be more more difficult to predict accurately. Furthermore, it is important to note that in all cases performance is imperfect - we are introducing additional measurement error so we need to verify that the sample enhancement is still worth it.

[Figure 5 about here.]

Translation. In our main results we allow our model to select between machine translations

[9] As we will show in Section 8, validation set performance is a good guide for true out of sample performance.

of our transcripts into English and a transliteration of the Bengali transcripts into Latin characters. In Appendix B.2 we explicitly compare the validation set performance of these representations along with a human translation of the interviews into English and the raw Bengali transcripts in their original Bengali script. We find that the machine English translation outperforms the human translation in most cases and in all cases the transliterated Bengali outperforms models trained on the raw Bengali transcripts.[10] This may be because in translating or transliterating the transcripts, we reduce some variance in the text while preserving the relevant content. Additionally, machine translations may be preferable to human translations because they will be more consistent across documents.

6 Assessing the Value of the Enhanced Sample

By enhancing our human annotated sample we increase the sample size, but introduce an additional source of measurement error. A priori, we therefore do not know if the enhanced sample has added value. Fortunately, we can assess the value of our enhanced sample once we have created it. By quantifying the measurement errors in our validation sets and comparing results in the human and enhanced sample we can assess whether our enhanced sample adds value along three dimensions: bias, efficiency and interpretability.

[Figure 6 about here.]

Bias. If our machine annotations introduce a sizeable bias, this is obviously a problem for any later analysis (e.g. if we always over-predict Secular aspirations for refugees we might get misleading results in the enhanced sample). We therefore need verify that any results in our enhanced sample are not driven by biases introduced in the predictions. In order to do this we test the association of prediction errors with household characteristics for each interview described in Table 1. We use the *validation set* predictions, and regress the implied prediction errors on a range of household characteristics. The F-statistic of these regressions tests whether there is evidence of a significant relationship between household characteristics and the predictions errors and so forms a natural test of bias.

[10]The average validation F1 scores across all annotations are 0.558 for Machine translation, 0.542 for Transliteration, 0.535 for Human translation and 0.420 for Raw Bengali.

Figure 6 shows the F-statistics for this bias test across each annotation. The test is carried about for each bootstrap iteration (show as the hollow points) as well as for the mean value across bootstraps (the solid points). The colour of each point indicates the significance level. A statistically significant F statistic here indicates that there may be a bias in the prediction errors that is related to the household characteristics.

In three cases there is evidence at the 5% level of a relationship between household characteristics and prediction errors - No Ambition, Education Neutral and Awareness Information Low, so we can look at these in more detail. The regressions in question, shown in Appendix C indicate that No Ambition, Education Neutral and Awareness Information Low are all under-predicted for refugees (i.e. the prediction errors are positive) and that Awareness Information Low is over-predicted for more educated parents. We will need to bear this in mind in our substantive analysis discussed in Section 7.

In addition to explicitly testing for bias, we also include a dummy variable for whether an interview is machine or human annotated in any regressions using enhanced sample data. This will account for any overall under or over prediction.

Efficiency. Even if measurement error does not introduce a bias in the machine annotations, it 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).[11]

To approximate the model error, we bootstrap the model by sampling the interviews with replacement B times. 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

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

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

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

out-of-sample predictions, but we find that the estimates are virtually identical in each.

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 out of sample, we only use the validation set predictions to compute these errors.[12] The estimated variance of this idiosyncratic error, $\hat{\sigma}^2_\varepsilon$ is then the variance of the validation set prediction errors. Of course, this variance has to be calculated the human sample, as these are the only observations for which we 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: $\hat{\sigma}^2_m = \hat{\sigma}^2_{\hat{y}} + \hat{\sigma}^2_\varepsilon$. 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.

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

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 $\hat{\sigma}^2_m > \hat{\sigma}^2_h$. Enhancing our sample therefore increases the number of observations but 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.

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

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 if a condition on the ratio of variances in the human

[12] As we show in Section 8, where we also compute errors for observations in a held-out test set, for a sufficiently large sample size, performance in the validation and in a held out test set coincide.

and machine annotated samples, relative to the increase in sample size, is met:

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

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 $N_h = 789$ and $N_m = 1618$, then our enhanced sample will have a smaller standard error for an estimate of the population mean if $\frac{\hat{\sigma}_m^2}{\hat{\sigma}_h^2} < 4$.

[Table 3 about here.]

Table 3 shows these variances computed at the interview level for the mean predictions across bootstraps in the cross-validated models. Standard errors for the population mean that have been adjusted as described above are also shown. We can see that in all cases the standard error of the population mean is lower than that of the human only sample. Enhancing the sample with our method thus increases the precision of these estimates, in spite of the the fact that predictive performance of our models is sometimes relatively low.

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.

Interpretability. We assess interpretability of our enhanced sample in two complimentary ways. Firstly, we compare the statistical significance of regressions of annotations on household characteristics in the enhanced and human annotated samples. Secondly, we use a supervised topic model trained on the predicted annotations.

[Figure 7 about here.]

Assuming text-based variables should be related to household characteristics, if our enhanced sample has improved the interpretability of our analysis it should give stronger evidence of a relation-

ship between the annotations and household characteristics.[13] We can therefore compare F statistics for regression of annotations on household characteristics in the human and enhanced samples. If the enhanced sample increases this F statistic relative to the human sample it suggests that the larger sample leads to more interpretable results in spite of the greater measurement error.

Figure 7 shows the F statistics of these regression in the human and enhanced sample. The F statistic in the human sample is shown as the cross and the enhanced sample as a hollow circle for each bootstrap iteration, with the solid circle for the mean prediction across bootstraps. In all cases the F statistic in the mean enhanced sample is higher than in the human sample. In some cases this difference is quite small though (e.g. Reliance on God) and in an individual bootstrap runs there is a decrease in some cases. There is thus no guarantee of increased interpretability when we enhance our sample, but in all our cases we see an increase for the mean across bootstrap iterations.

An alternative sense of interpretability relates to the relationship between the predicted annotations and the text itself. The classifier models we use to create our enhanced sample are in general optimised for prediction rather to give directly interpretable relationships between the text and annotations. However, once we have these predictions we can use an alternative model to assess which text features are associated with an annotation in a more interpretable way. We thus estimate a supervised topic model Blei and McAuliffe (2008) for our machine prediction of each annotation, based on the interview text. We can thus verify that the topics most (and least) associate with the predictions for each annotations roughly correspond to our definitions of that annotation.

[Figure 8 about here.]

Figure 8 shows the output of these supervised topic models for the two aspiration annotations. There are ten topics, represented along the vertical axes by the ten most highly weighted words in each topic. Each topic is then associated with a coefficient where a positive coefficient means that topic is more likely to be associates with that annotation. We can thereby verify that the text features associated with the predictions for each annotation correspond to our understanding of each annotation. In this case, we see that the topic most associated with secular aspirations highly weights

[13] Note that this test does not require any assumptions on how the text and household characteristics are related, just the relatively weak assumption that there is some relationship between them.

words such as “good”, “dream”, “human” and “educated”, consistent with our definition of secular aspirations. Similarly, religious aspirations are associated with topics that place a high weight on religious terms like “hafez” and “madrasa”, “god” and “allah”.

7 Results

The overarching trend is that the results with the enhanced sample have smaller standard errors. For instance, if we compare the correlation matrix computed on the enhanced sample in Figure 9, with that on the human sample (Appendix C); we see that the enhanced sample shows a much higher proportion of statistically significant correlations. Crucially, the signs of correlations and coefficients across the human and enhanced samples are the same. Our method thus appears to be successful in increasing the available sample size but does not introducing a bias that changes interpretation.

[Figure 9 about here.]

Focusing on Figure 9, we first look at the correlations within each of the three code domains - Ambition, Aspirations, and Capacity. Within the Ambition domain, having “no ambition” is negatively correlated with wanting a high education, a salaried job or being an entrepreneur, but positively correlated with marriage. Ambition for a high education is on the hand negatively correlated with marriage but positively with salaried employment. It is interesting to note also that marriage is negatively correlated with salaried employment suggesting that parents who are focused on getting a child married are less likely to say that they want her to have a salaried job. Note that parents who want their children to have a religious education tend also talk about wanting them to have higher levels of secular education.

Within the Aspiration domain, however, parents who profess to have Secular Aspirations for their children do not say that they have Religious Aspirations suggesting that aspirations, in the sense that Agnes Callard defines it, is capturing something different from ”ambition.”

The codes in the Capacity domain - which attempt to capture Arjun Appadurai’s concept of Navigational Capacity - tend to move in the same direction. People who display High Ability or good navigational capacity also tend to have higher budgets and more information. Conversely people who

display Low Ability tend to have lower budgets, but are positively correlated with both Low and High informational awareness suggesting that Low Ability is not necessarily a function of low information. Similarly people with low information seem to report both high and low budgets.

Looking at the correlations across domains, the codes in the Capacity domain are generally positively correlated with high ambition. People who show High Ability also have higher education ambition, and more likely to want a salaried job for their children. Similarly people who have higher budgets and high information awareness also more likely to have higher education ambitions and want salaried jobs for their children. On the other hand, parents who report that they are budget constrained tend to want a religious education for their children, and want them to be married, have vocational training and entrepreneurial jobs.

In the Aspirations domain, parents who report having Secular Aspirations for their children also want higher levels of education and salaried jobs for their children, and tend to be of higher ability and have high information awareness. Parents with Religious Aspirations for their children show a positively correlation both with higher levels of Religious and Secular education but also report a higher Reliance on God.

7.1 Ambition

Next, in Tables 4-9 we report results from a set of reduced form regressions where we regress the codes against household characteristics from the 2019 "baseline" survey. We include refugee status, number of children, whether it is a female headed household, the age of the head, the parent's years of education, whether s/he is religiously educated, whether the child is female, the household's 2019 asset index, its 2019 income, and a trauma score for the head of the household. We report results from the human-coded sub-sample and the enhanced sample next to each other for all the regressions and again note that the enhanced sample regressions tend to be more precisely estimated.[14] We also control for whether the data is from the second qualitative round (which is Round 3 because the baseline survey was Round 1), and in the enhanced sample regressions we include a dummy for whether the household interview was human annotated.

[14] Across all regression tables, results using the human annotated sample are reported in odd numbered columns, and results using the enhanced sample in even numbered columns.

[Table 4 about here.]

We start with the Ambition domain, reporting Education results in Table 4. Parents with higher levels of education have higher education ambitions for their children, and are less likely to want a religious education. Wealthier parents with more household assets also report higher education ambition. However, parents report lower education ambition for their female children. Parents also report lower Religious Education ambitions for female children, as do more educated parents. Parents with a religious education, however, also report wanting their children to be religiously educated.

[Table 5 about here.]

Table 5, which reports results on Employment ambition, shows that parents of female children and less likely to want report wanting them to have salaried employment or to be entrepreneurs. More educated parents are more likely to want their children to have salaried employment. Religiously educated parents are more likely to want their children to have vocational training and less likely to want their children to be entrepreneurs.

[Table 6 about here.]

Table 6 reports results from parents who report "no ambition" and ambitions for marriage and migration. Refugees are more likely to not have ambitions for their children, and female heads of household are less likely to do so. We should bear in mind here that our bias tests found that the "no ambition" code was systematically under predicted for refugees in our machine predictions (Table 23). The coefficient on refugee status in the enhanced sample is thus likely an underestimate, although it remains positive and significant. Parents with a female eldest child are much more likely to have marriage oriented ambitions, and less likely to speak about migration.

7.2 Aspirations

Table 7 reports results on Aspirations - coded into either Secular or Religious. Interestingly parents are less likely to report having aspirations, both secular and religious, for female children. More educated parents tend to report more secular aspirations, as do younger parents. While parents with a religious education are more likely to have religious aspirations for their children.

[Table 7 about here.]

7.3 Navigational Capacity

Moving to the codes categorized under Navigational Capacity. We first look at Low and High Ability and Low and High Budget in Table 8. Note that refugees are less likely to be coded as having low ability. This is likely because of the selection process associated with having escaped war and conflict and having reached the relative safety of the camp which would make surviving refugees people who are more capable than average. Ability Low is also less likely to be present with more educated parents, and those from wealthier households. Ability High shows consistent results with more educated parents more likely to be those of high ability, while older parents and female headed households are less likely to demonstrate high levels of ability. Refugees, understandably, are more likely to talk about being constrained by budgets as are less wealthy households, less educated educated parents and parents with larger numbers of children. More educated parents, consistently, are more likely to report having relatively high budgets.

[Table 8 about here.]

Table 9, reports results from the other Ability codes - information awareness, reliance on God and "vague-non-specific answers." Refugees are less likely to give vague non-specific answers to the navigational capacity question, and are also less likely to report a reliance on God as a response to the question. Not surprisingly, better educated parents are also likely to have high information awareness and less likely to report a reliance on God. Parents of girls more likely to show low informational awareness and less likely to show high informational awareness.

[Table 9 about here.]

In summary, the three domains of codes for aspirations, Ambition, Aspiration and Ability show some interesting patterns. First, they seem to be distinct concepts with different determinants. Callard's distinction between Ambition and Aspiration is important, and Appadurai's notion of Navigational Capacity also matters - just having an ambition or aspiration for your child is not enough, there is

a lot of heterogeneity in the capacity of parents to know how to achieve these goals. It is affected by their ability to articulate a clear strategy of how to get there, but also by constrained budgets and information. There is clear evidence of gender bias - parents of girls tend to have less education and job ambitions for them and seem to be more focused on marriage. There is some evidence that refugees tend to be of higher ability than average, but more constrained by budgets. Finally, it is clear that parents have both religious and secular aspirations for their children that seem to be distinct from each other - with religious aspirations more likely to be professed by parents who have a religious education.

7.4 Does Qualitative Add Value?

Aspirations related to education goals and ambitions have been extensively studied by economists, which raises the question of whether the open-ended questions on ambition add value. To answer this we included a standard structured question on education ambition in round 2, in addition to the open-ended questions. We regressed the same set of exogenous variables on the quantitative education response, and then added the qualitative ambition codes to the regression - results are reported in Table 10.

[Table 10 about here.]

The quant and qual education ambition results have very similar signs and significance levels, but the qual question seems to add nuance and capture additional variation. The refugee coefficient on quant education ambition is strongly negative which would lead us to to believe that refugees have much lower education ambitions than hosts. The qualitative code regressions (reported in Table 6), however, reveal a more complex interpretation. Refugees have similar levels of high and low education ambition as hosts. However, they are much more likely to report "Neutral" Education Ambition. Neutral is the code we used for situations where respondents expressed helplessness in context of ambitions or said that were unable to have dreams or plans on a given topic. This tells us that it is not that refugees are less ambitious on education than hosts, but that having had disrupted lives they have more trouble expressing a clear education ambition. Table 12 also shows that when the

qual education ambition codes are added to the reduced form regression on quant education ambition, they add explanatory power without substantially changing the coefficients on the original set of right hand side variables suggesting that are capturing additional variation.

While ambition is relatively easy to turn into a structured question, latent concepts like "aspiration" and "capacity" are harder because they are more subtle. Since we conducted these surveys in the COVID period we were unable to conduct the extensive field work needed to develop and pretest good structured questions on (Callardian) aspirations and navigational capacity to enable a direct comparison between quant and qual versions of these concepts. However, we show that asking relatively straightforward open-ended questions on aspirations and navigational capacity captured a great deal of content. This raises the question of whether (a) developing a quantitative module would have added value to these intrinsically more subtle latent concepts, and (b) whether relying entirely on quantitative representations could detract from understanding the point of view of respondents.

8 Varying sample sizes

A key question for researchers looking to scale up manual annotations using supervised models is how many of their documents to annotate in order to make the most of their data. In this Section, we vary the size of both the human annotated (N_h) and machine annotated (N_m) samples to explore this question. We find that, while out-of-sample performance and interpretability of results increases with the number of human annotated interviews, both of these display diminishing returns to scale. Results remain largely the same if at least 400 interviews are annotated. While these results are encouraging, further work will be needed to test whether similar results will be found in other datasets.

We also show in a simple cost-benefit-analysis exercises based on either (i) maximising the average F statistic in enhanced sample regressions or (ii) minimising the standard errors of key estimated coefficients, recommend a role for machine annotation.

8.1 Performance

Increasing N_h will improves the out-of-sample performance of the classifier models as they can be trained on a larger sample. However, annotating extra interviews without compromising the anno-

tation quality is both time consuming and expensive. It is therefore useful to explore how much performance improves as the annotated sample size N_h increase. To assess this we draw subsets of our human annotated sample without replacement, varying the sample size from 100 to the full 789 (drawing a 10 separate samples for each possible N_h). We then train our classifier models on these subsets, selecting the text representation and model independently in each case. Given that we now no longer use the entire human annotated set for training, we can also quantify performance in the out-of-bag test set as well as the validation set performance of each model.

As N_h increases, there are improvements in both validation set and out-of-bag test set performance. When the human annotated sample is very small, out-of-bag test set performance is worse than validation set performance, however once we have 400 annotations, validation set and test set performance are pretty much the same for most annotations. Averaging across annotations, moving from 100 to 200 human annotated interviews increases the out of sample F1 score by 0.05, moving to 300 then gives an extra 0.03, moving to 400 an extra 0.013. So while increasing the size of the human annotated set does improve prediction performance, there do appear to be diminishing returns here.[15] The validation and test set performance for each annotation as N_h varies is shown in Appendix D. vary with N_h for each annotation. Whether our machine annotations introduce bias or increase efficiency depends on the out-of-sample performance of the supervised models. In line with the results on performance, we find that a larger human annotated sample reduces bias and decreases measurement error, but that this also displays diminishing returns after around 400 or 500 annotated interviews.

8.2 Interpretability

Interpretability can be affected by increasing both N_h and N_m. In other words, we may be able to get stronger results by either annotating more interviews or by conducting additional interviews and machine annotating them. For each of the models trained on samples from 100 to the full 789 human annotated interviews we therefore generate predictions for a randomly sampled subset of the unannotated interviews, from 200 to 1,400 (again drawing a 10 separate samples for each possible

[15]There may of course be non-linearities in the performance of the text based classifier models, particularly if for very large samples a more sophisticated model becomes feasible. However, for the number documents that it would be realistic to annotate manually this is unlikely to be a concern.

combination of N_h and N_m).

Our enhanced sample interpretability measure - the F statistic of a regression of the annotation on household characteristic - generally increases as extra interviews are added through N_h (on the horizontal axis) and N_m (on the vertical axis). Averaging across all annotations, we find adding 100 more human annotated interviews increases the F statistic in the enhanced sample by 8.4% while adding 100 more machine annotated interviews increases it by an average of 6.1%. Annotating an additional 100 existing interviews by hand therefore increases the F statistic by around 2.3%. On average, an additional interview therefore has around 3 times the benefit of annotating an existing interview. This interpretability measure for each annotation as both N_h and N_m vary is shown in Appendix D.

[Figure 10 about here.]

We can also look at the effect of increasing N_h while holding $N = N_h + N_m$ fixed. Intuitively, we can think of this as adding human annotations to some of the existing interviews that are currently machine annotated. Figure 10 shows show the F statistic test for interpretability changes for each annotation as N_h is increased while N is constant. In blue we see the F statistic for the human only sample - a higher N_h will of course increases this as it increases the size of the human sample and so we get a more statistically significant relationship between the text and household characteristics. In green we see the F statistic on the enhanced sample. While a higher N_h does generally increase the F statistic in the enhanced sample (as predictions are more accurate) the overall sample size doesn't increase.

Interestingly, while the enhanced sample F statistics in Figure 10 do increase somewhat with N_h, this increase is relatively small suggesting that it may be sufficient to annotate a relatively small number of interviews. In many cases there appear to be diminishing returns to extra annotations - at some point the enhanced sample is good enough that adding additional annotations doesn't really make a difference.

As an alternative to the very general approach of focusing on the F-statistic across all annotations, we can instead focus on how estimates of specific coefficients change as the sample sizes change. This approach may be more appropriate in many applications where there will be a specific effect or effects

that are of primary interest. To illustrate this, Figure 11 shows how two coefficients of interest from Section 7 (the coefficient on refugee status for ability low and the coefficient on Female eldest child in Secular Aspirations) vary as the size of the human annotated and machine annotated samples changes. This distribution of the coefficients in the enhanced sample are shown in red and the human sample in blue. Rather strikingly, the enhanced sample coefficients do not change very much with N_h, again suggesting that enhancing the sample is useful even with a very small number of human annotated interviews.

[Figure 11 about here.]

8.3 A cost benefit analysis

To give a sense of the trade-offs a researcher may face in deciding on how many interviews to conduct and how many to annotate we conduct a simple cost benefit analysis exercise based on the results discussed above. In our case, the marginal cost of conducting a single additional interview was around $12 while the marginal cost of annotating one additional interview was around $3 (all costs here are given in 2021 US dollars).[16] For a given budget (between $10,000 and $20,000) we then find the combination of N_h and N_m that maximises some objective. We report results under three different objectives here:

1. Maximising the average F statistic in the enhanced samples for a regression of annotation on household characteristics, across all annotations (i.e. Figure 7).

2. Minimise the 95% confidence interval of the refugee status coefficient in the enhanced sample regression on ability low (i.e. upper panels in Figure 11).

3. Minimise the 95% confidence interval of the refugee status coefficient in the enhanced sample regression on ability low (i.e. upper panels in Figure 11).

For each objective, increasing either N_h or N_m will generally lead to a better outcome, but will be more expensive. In our case, disregarding unannotated interviews for which we have missing data on

[16] This cost figure for annotation is likely a substantial underestimate as the major difficulty with high quality annotation is finding annotators with the adequate skills.

household characteristics, we conducted 2,270 interviews of which 789 where annotated. Ignoring fixed costs, this had an estimated cost of $30,000.

[Figure 12 about here.]

For a given budget and starting point, we can thus calculate out the optimal mix of N_m and N_h. Figure 12 shows how different combinations of N_h and N_m perform across the three objectives we consider. These points can be thought of as forming iso-cost curves: for a given budget we can choose the allocation across N_h and N_m that maximises our objective. Unsurprisingly, this curve is considerably smoother when the average enhanced sample log F statistic is the objective as this encompasses all annotations and household characteristics. In contrast, when the objective is a single coefficient there is likely a lot more idiosyncratic noise in the performance for a given combination of N_h and N_m.

[Table 11 about here.]

Table 11 shows the optimal combination of N_h and N_m for budgets of $10,000, $15,000 and $20,000. In each case, a mix of human annotated and machine annotated interviews appears to be preferred, suggesting that there is value in the enhanced sample procedure set out in this paper. Interestingly, in the case of the most general objective (the average F statistic), our exercise suggests that 500 interviews be human annotated and then any extra budget be used for machine annotated interviews. This further suggests that a relatively small human annotated sample can lead to good results when combined with machine annotation of a larger sample.

9 Conclusion

Interpretative qualitative analysis, which is a common tool in anthropology, sociology and related disciplines, is not used by economists but is potentially of considerable value. It is predicated on a close, careful, inductive, and nuanced human reading and coding of textual information – usually on open-ended interviews with respondents. The method is "reflexive" in that it allows for data to be collected and analyzed in a more bottom-up manner that is driven more by respondents than by

researchers. Instead of requiring respondents to provide quantitative responses to questions that may force false precision on a latent concept, it allows respondents to speak about a topic in a manner that is closer to how they understand the concept resulting in more accurate and nuanced responses.

Interpretative qualitative analysis could be potentially very useful for a variety of topics of interest to contemporary economists such as well-being, cultural change, social norms, networks, decision-making and the topic of this paper – aspirations. The could also be potentially of great value in understanding change processes and mechanisms in experiments and randomized trials. However, the high level of human effort required in employing the method has generally restricted its use to small samples. This is one reason why it has not been used by economists. This paper present a machine learning method to extend interpretative human coding (iQual) to large, representative samples. The method takes a smaller sub-sample which is coded by human, interpretative, coders. This human-coded sub-sample is used as a training set to use extend to a larger, representative "enhanced" sample that allows us to standard econometric tools to analyze the data. Rather than recommend a particular text representation or supervised model, we select both to optimise for out-of-sample predictive performance. We demonstrate that this sample enhancement adds value by testing for bias and showing that the enhanced sample increases the efficiency and interpretability of analysis.

We apply the method to over 2,000 open-ended interviews on parent's aspirations for children collected from a representative sample of Rohingya refugees and their Bangladeshi hosts in Cox's Bazaar, Bangladesh. In the open-ended interviews we are able to show that aspirations have dimensions that are much broader than how they have generally understood in the economics literature. In economics aspirations have generally been views as "ambition" – specific goals on education or jobs that parents have for their children. We show that open-ended interviews allow this to be broadened to include the moral and spiritual dimensions of aspirations – for instance being a "good person" or a "good Muslim," and to understand a parent's "navigational capacity" - their ability to act in a way that allows aspirations to be realized. We show that these three distinct domains of "ambition," "aspiration" and "navigational capacity" are correlated in interesting ways with each other, and have distinct relationships with exogenous household characteristics.

We explore the role of the size of the human annotated sample for the value of the sample en-

hancement through a series of simulations. We find that, in our application at least, annotating even a relatively small number of interviews and scaling these up with NLP can be a cost effective way of analysing a large corpus of open ended interviews. These simulations also demonstrate the robustness of our results to different annotated sets.

This paper comes with a Python package, *iQual* (https://github.com/worldbank/iQual) that implements the supervised models and various tests that we perform.

Author affiliations.

Julian Ashwin: Department of Economics, London Business School.

Vijayendra Rao: World Bank Group.

Monica Biradavolu: Qual Analytics.

Aditya Chhabra: World Bank Group.

Arshia Haque: World Bank Group.

Afsana Khan: Princeton School of Public and International Affairs, Princeton University.

Nandini Krishnan: World Bank Group.

A Qualitative coding

A.1 Coding Tree

[Table 12 about here.]

[Table 13 about here.]

[Table 14 about here.]

A.2 Achieving Cross-coder Agreement

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 789 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 15 illustrates the process by which codes were refined to be more nuanced and context-specific as a result of the review process. As an example take 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 a madrassa or noorani school, these instances were coded as Religious Aspiration. After review and seeking expert input, we understood that these references should 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 ranked codes: Ambition:Education:High, Ambition:Education:Neutral or Ambition:Education:Low. As a result, the definitions for both the aspirations and the ambition group of codes were better specified, leading to a deeper understanding of respondents' hopes and dreams for their children.

[Table 15 about here.]

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 that code would be refined to reach a clearer definition.

In the example shown in Table 16, MB rated the code "Salaried Employment" as fuzzy as she observed religious jobs such as "madrassa teacher" coded under salaried employment by both coders. This was resolved by further refining the "Salaried Employment" code and creating further sub-codes 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 16 about here.]

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".

B Modelling

B.1 Model details

[Table 17 about here.]

[Table 18 about here.]

[Table 19 about here.]

[Table 20 about here.]

[Table 21 about here.]

[Table 22 about here.]

B.2 Translation methodology

[Figure 13 about here.]

C Additional Results

[Table 23 about here.]

[Table 24 about here.]

[Figure 14 about here.]

D Varying Sample Size

[Figure 15 about here.]

[Figure 16 about here.]

References

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019), Optuna: A next-generation hyperparameter optimization framework, *in* 'Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining', pp. 2623–2631.

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.

Andre, P., Haaland, I., Roth, C. and Wohlfart, J. (2021), 'Narratives about the macroeconomy'.

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

Appadurai, A. (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* pp. 59–84.

Armborst, A. (2017), 'Thematic proximity in content analysis', *Sage Open* **7**(2), 2158244017707797.

Ashwin, J., Kalamara, E. and Saiz, L. (2021), 'Nowcasting euro area gdp with news sentiment: a tale of two crises'.

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', pp. 121–128.

Bonikowski, B. and DiMaggio, P. (2022), 'Mapping culture with latent class analysis: A response to eger and hjerm', *Nations and Nationalism* **28**(1), 353–365.

Bonikowski, B. and Nelson, L. K. (2022), 'From ends to means: The promise of computational text analysis for theoretically driven sociological research', *Sociological Methods & Research* **51**(4), 1469–1483.

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.

Ferguson-Cradler, G. (2021), 'Narrative and computational text analysis in business and economic history', *Scandinavian Economic History Review* pp. 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, World Bank.

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.

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 21-104, 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', pp. 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.

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.

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.

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.

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.

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.

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', IEEE, pp. 1–10.

List of Tables

Table 1: Quantitative variable summary statistics

Statistic	N	Mean	St. Dev.	Notes
Refugee status	2,407	0.460	0.498	Dummy variable, 1 for refugee
Female eldest child	2,294	0.525	0.499	Dummy variable, 1 for female
Eldest child's age	2,295	10.069	4.963	Integer
Female household head	2,407	0.185	0.388	Dummy variable
Household head's age	2,287	33.231	10.425	Integer
Number of children	2,405	2.675	1.463	Integer
Parent's years of education	2,398	3.548	3.837	Integer
Parent's religious education	2,398	0.036	0.186	Dummy variable
Asset Index	2,406	0.147	1.820	Principle Component of assets owned following Filmer and Pritchett (2001)
Household Income	2,407	1.125	2.340	Income for last month in 10,000s Bangladeshi taka
Trauma Event Score	2,287	2.641	2.410	Sum of positive responses for experience of twelve possible traumatic events following Harvard Trauma Questionnaire
Quant Education Ambition	1,267	4.272	1.657	Ordered categorical (1-7) from question on parents' ambitions for eldest child's education

Table 2: Qualitative human annotations summary

Category	Annotation	Proportion of QA pairs R2	R3	Total	Proportion of interviews R2	R3	Total
Aspirations	Religious	0.036	0.028	0.030	0.208	0.234	0.230
	Secular	0.053	0.028	0.036	0.332	0.332	0.360
Ambition	No Ambition	0.021	0.013	0.016	0.129	0.129	0.137
	Salaried Employment	0.094	0.101	0.099	0.310	0.362	0.354
	Vocational Training	0.007	0.004	0.005	0.041	0.041	0.045
	Entrepreneur	0.031	0.012	0.018	0.154	0.122	0.150
	Education Low	0.013	0.035	0.028	0.094	0.475	0.312
	Education Neutral	0.185	0.062	0.101	0.772	0.574	0.691
	Education High	0.064	0.048	0.053	0.375	0.454	0.427
	Education Religious	0.035	0.016	0.022	0.210	0.168	0.198
	Marriage	0.082	0.036	0.050	0.385	0.396	0.418
	Migration	0.022	0.007	0.012	0.104	0.079	0.097
Capacity	Vague Non-Specific	0.066	0.017	0.033	0.420	0.234	0.349
	Reliance on God	0.039	0.017	0.024	0.243	0.228	0.253
	Ability High	0.048	0.032	0.037	0.311	0.424	0.391
	Ability Low	0.035	0.033	0.034	0.230	0.360	0.321
	Budget High	0.033	0.013	0.020	0.200	0.188	0.212
	Budget Low	0.111	0.039	0.062	0.522	0.401	0.492
	Awareness Information High	0.070	0.031	0.043	0.387	0.297	0.367
	Awareness Information Low	0.007	0.012	0.010	0.061	0.145	0.114
Other	Covid Impacts	0.022	0.021	0.021	0.149	0.272	0.226
	Public Assistance	0.020	0.005	0.010	0.137	0.058	0.107
	Worries Anxieties	0.049	0.014	0.025	0.268	0.175	0.242

Table 3: Measurement error variances

Category	Annotation	$\hat{\sigma}_h^2$	$\hat{\sigma}_{\hat{y}}^2$	$\hat{\sigma}_{\varepsilon}^2$	$\hat{se}_h$	$\hat{se}_{enh}$
Aspirations	Religious	0.0060	0.0073	0.0020	0.0027	0.0018
	Aspirations Secular	0.0090	0.0084	0.0042	0.0034	0.0022
Ambition	No Ambition	0.0015	0.0010	0.0010	0.0014	0.0009
	Salaried Employment	0.0156	0.0175	0.0055	0.0045	0.0029
	Vocational Training	0.0014	0.0010	0.0003	0.0013	0.0007
	Entrepreneur	0.0053	0.0075	0.0015	0.0026	0.0018
	Education High	0.0093	0.0090	0.0055	0.0034	0.0023
	Education Neutral	0.0245	0.0267	0.0108	0.0056	0.0037
	Education Low	0.0027	0.0023	0.0014	0.0019	0.0012
	Education Religious	0.0047	0.0049	0.0023	0.0024	0.0016
	Marriage	0.0133	0.0127	0.0016	0.0041	0.0024
	Migration	0.0042	0.0026	0.0007	0.0023	0.0012
Capacity	Vague Non-Specific	0.0062	0.0073	0.0049	0.0028	0.0021
	Reliance on God	0.0041	0.0043	0.0020	0.0023	0.0015
	Ability High	0.0064	0.0098	0.0036	0.0029	0.0021
	Ability Low	0.0057	0.0050	0.0038	0.0027	0.0018
	Budget High	0.0046	0.0055	0.0025	0.0024	0.0017
	Budget Low	0.0156	0.0116	0.0060	0.0044	0.0026
	Awareness Information High	0.0091	0.0096	0.0070	0.0034	0.0024
	Awareness Information Low	0.0010	0.0008	0.0010	0.0011	0.0008

Note: This Table reports the measurement error variances and standard error on the sample mean for each annotation. The $\hat{\sigma}_h^2$ column reports the variance of the annotation in the human sample. The $\hat{\sigma}_{\hat{y}}^2$ column reports the variance of the machine annotated sample, across all bootstraps. The $\hat{\sigma}_{\varepsilon}^2$ column reports the variance of validation set errors. Finally, $\hat{se}_h$ and $\hat{se}_{enh}$ represent the standard errors of the sample mean in the human and enhanced samples respectively, taking the measurement error adjustments into account.

Table 4: Educational ambition variables and household characteristics

	Dependent variable:							
	Education High		Education Neutral		Education Low		Education Religious	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
R3	−0.029*	−0.021**	−0.122***	−0.169***	0.037***	0.040***	−0.005	0.0004
	(0.016)	(0.009)	(0.023)	(0.012)	(0.009)	(0.004)	(0.012)	(0.006)
Machine annotated		0.001		0.027***		−0.004**		0.0003
		(0.004)		(0.006)		(0.002)		(0.003)
Refugee	−0.005	−0.008	0.032**	0.026***	0.005	0.004	0.012	0.003
	(0.011)	(0.006)	(0.016)	(0.008)	(0.006)	(0.003)	(0.008)	(0.004)
Number of children	−0.0004	−0.001	−0.005	−0.004*	−0.002	−0.001	−0.003	−0.001
	(0.003)	(0.001)	(0.004)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Female HH head	−0.002	0.003	0.008	−0.003	−0.004	0.001	−0.011	−0.001
	(0.009)	(0.005)	(0.014)	(0.007)	(0.005)	(0.002)	(0.007)	(0.004)
Age of HH head	0.0001	0.00001	−0.0002	−0.001***	−0.0002	0.00005	0.0004	0.0001
	(0.0004)	(0.0002)	(0.001)	(0.0003)	(0.0002)	(0.0001)	(0.0003)	(0.0002)
Parent's years of education	0.005***	0.004***	−0.002	0.001	−0.001*	−0.001**	−0.001	−0.001*
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.0003)	(0.001)	(0.0004)
Religiously educated parent	0.023	0.004	−0.020	0.001	−0.008	−0.006	0.008	0.018**
	(0.018)	(0.011)	(0.026)	(0.015)	(0.010)	(0.005)	(0.013)	(0.008)
Female eldest child	−0.006	−0.013***	−0.008	0.007	0.001	−0.001	−0.011**	−0.014***
	(0.007)	(0.004)	(0.011)	(0.006)	(0.004)	(0.002)	(0.005)	(0.003)
Age of eldest child	0.0004	0.0001	−0.001	0.0002	−0.0003	−0.0004**	−0.001*	−0.001***
	(0.001)	(0.0003)	(0.001)	(0.0004)	(0.0003)	(0.0001)	(0.0004)	(0.0002)
HH asset index	0.003	0.005***	0.002	0.001	−0.001	−0.002**	0.001	0.001
	(0.003)	(0.002)	(0.005)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
HH income	0.002	−0.001	0.003	−0.0001	−0.001	−0.0001	−0.002	−0.001
	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)	(0.0005)	(0.002)	(0.001)
Parent trauma experience	−0.002	0.001	−0.001	0.0003	0.0003	0.001	0.0003	0.001
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.0004)	(0.001)	(0.001)
Constant	0.054***	0.069***	0.243***	0.247***	0.029***	0.014***	0.047***	0.054***
	(0.017)	(0.010)	(0.025)	(0.013)	(0.009)	(0.004)	(0.012)	(0.007)
Observations	696	2,177	696	2,177	696	2,177	696	2,177
R^2	0.086	0.070	0.251	0.297	0.104	0.149	0.060	0.050
F Statistic	5.347***	12.481***	19.106***	70.149***	6.591***	29.047***	3.662***	8.842***

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

Table 5: Employment ambition variables and household characteristics

	Dependent variable:					
	Salaried Employment		Vocational Training		Entrepreneur	
	(1)	(2)	(3)	(4)	(5)	(6)
R3	0.021	0.014	−0.005	−0.002	−0.013	−0.028***
	(0.021)	(0.012)	(0.006)	(0.003)	(0.011)	(0.007)
Machine annotated		−0.002		−0.002		0.007**
		(0.006)		(0.001)		(0.003)
Refugee	−0.014	−0.013*	−0.002	0.0003	−0.009	−0.001
	(0.014)	(0.008)	(0.004)	(0.002)	(0.007)	(0.005)
Number of children	0.001	−0.003*	−0.001	0.0005	0.003	0.002*
	(0.003)	(0.002)	(0.001)	(0.0004)	(0.002)	(0.001)
Female HH head	−0.026**	−0.001	−0.001	−0.002	0.014**	0.001
	(0.012)	(0.007)	(0.004)	(0.002)	(0.007)	(0.004)
Age of HH head	0.0001	−0.00003	−0.0001	−0.00002	0.0004	0.0002
	(0.001)	(0.0003)	(0.0002)	(0.0001)	(0.0003)	(0.0002)
Parent's years of education	0.006***	0.004***	−0.001	−0.0002	−0.0004	−0.001*
	(0.001)	(0.001)	(0.0004)	(0.0002)	(0.001)	(0.0005)
Religiously educated parent	0.008	−0.009	0.010	0.007**	−0.007	−0.021**
	(0.023)	(0.014)	(0.007)	(0.003)	(0.013)	(0.009)
Female eldest child	−0.018*	−0.012**	0.004	0.003**	−0.012**	−0.013***
	(0.009)	(0.006)	(0.003)	(0.001)	(0.005)	(0.003)
Age of eldest child	−0.0002	−0.0003	0.0001	0.00005	−0.00005	0.0003
	(0.001)	(0.0004)	(0.0002)	(0.0001)	(0.0004)	(0.0003)
HH asset index	0.005	0.003	0.0002	0.0001	0.001	−0.001
	(0.004)	(0.002)	(0.001)	(0.0005)	(0.002)	(0.001)
HH income	−0.0003	0.002	0.0003	0.00001	−0.001	−0.001
	(0.003)	(0.001)	(0.001)	(0.0003)	(0.002)	(0.001)
Parent trauma experience	0.0001	0.002	−0.001	−0.0003	−0.001	−0.0001
	(0.002)	(0.001)	(0.001)	(0.0003)	(0.001)	(0.001)
Constant	0.095***	0.112***	0.014**	0.005*	0.017	0.026***
	(0.022)	(0.013)	(0.007)	(0.003)	(0.012)	(0.008)
Observations	696	2,177	696	2,177	696	2,177
R^2	0.094	0.050	0.015	0.010	0.042	0.040
F Statistic	5.883***	8.777***	0.874	1.662*	2.467***	6.927***

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

Table 6: Other ambition variables and household characteristics

	Dependent variable:					
	No Ambition		Marriage		Migration	
	(1)	(2)	(3)	(4)	(5)	(6)
R3	0.004	0.003	−0.086***	−0.093***	−0.013	−0.001
	(0.007)	(0.003)	(0.018)	(0.010)	(0.008)	(0.005)
Machine annotated		−0.005***		−0.006		0.002
		(0.001)		(0.005)		(0.002)
Refugee	0.014***	0.005***	0.007	0.007	0.007	0.002
	(0.004)	(0.002)	(0.012)	(0.006)	(0.006)	(0.003)
Number of children	−0.001	−0.0003	0.004	0.003*	0.002	0.0001
	(0.001)	(0.0004)	(0.003)	(0.002)	(0.001)	(0.001)
Female HH head	−0.008**	−0.002	0.007	−0.0002	−0.003	0.001
	(0.004)	(0.002)	(0.011)	(0.006)	(0.005)	(0.003)
Age of HH head	0.0003*	0.0001	−0.0004	−0.0004	−0.0001	0.0002
	(0.0002)	(0.0001)	(0.0005)	(0.0002)	(0.0002)	(0.0001)
Parent's years of education	−0.00002	−0.0001	−0.001	0.0002	0.00000	−0.001*
	(0.0004)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.0003)
Religiously educated parent	−0.006	−0.001	−0.008	−0.013	0.006	−0.001
	(0.008)	(0.003)	(0.020)	(0.012)	(0.010)	(0.006)
Female eldest child	0.001	0.001	0.033***	0.045***	−0.011***	−0.008***
	(0.003)	(0.001)	(0.008)	(0.005)	(0.004)	(0.002)
Age of eldest child	0.00005	0.0001	0.002**	0.002***	0.0001	−0.0002
	(0.0002)	(0.0001)	(0.001)	(0.0004)	(0.0003)	(0.0002)
HH asset index	−0.002	−0.001	−0.007*	−0.003	0.0004	−0.0004
	(0.001)	(0.0005)	(0.003)	(0.002)	(0.002)	(0.001)
HH income	−0.00000	−0.0001	0.004*	−0.0003	−0.001	−0.00003
	(0.001)	(0.0003)	(0.002)	(0.001)	(0.001)	(0.001)
Parent trauma experience	0.0001	−0.0002	0.001	−0.001	−0.001	−0.001
	(0.001)	(0.0003)	(0.002)	(0.001)	(0.001)	(0.0005)
Constant	−0.010	0.005*	0.049***	0.053***	0.021**	0.018***
	(0.007)	(0.003)	(0.019)	(0.011)	(0.009)	(0.005)
Observations	696	2,177	696	2,177	696	2,177
R^2	0.064	0.030	0.098	0.105	0.034	0.015
F Statistic	3.888***	5.211***	6.199***	19.488***	2.025**	2.619***

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

Table 7: Aspiration variables and household characteristics

	Dependent variable:			
	Secular		Religious	
	(1)	(2)	(3)	(4)
R3	−0.035**	−0.029***	0.016	0.012
	(0.015)	(0.008)	(0.013)	(0.008)
Machine annotated		0.008**		0.006*
		(0.004)		(0.004)
Refugee	−0.005	−0.005	−0.00002	−0.002
	(0.010)	(0.005)	(0.009)	(0.005)
Number of children	0.002	−0.001	−0.002	−0.001
	(0.003)	(0.001)	(0.002)	(0.001)
Female HH head	0.002	−0.002	−0.012	0.001
	(0.009)	(0.005)	(0.008)	(0.005)
Age of HH head	−0.0005	−0.0004**	0.001*	0.0002
	(0.0004)	(0.0002)	(0.0003)	(0.0002)
Parent's years of education	0.002*	0.002***	−0.0002	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Religiously educated parent	0.003	0.008	0.015	0.017*
	(0.017)	(0.010)	(0.015)	(0.010)
Female eldest child	−0.013*	−0.012***	−0.021***	−0.022***
	(0.007)	(0.004)	(0.006)	(0.004)
Age of eldest child	0.0002	0.00003	−0.001**	−0.001***
	(0.001)	(0.0003)	(0.0005)	(0.0003)
HH asset index	−0.001	−0.001	−0.005*	−0.002
	(0.003)	(0.001)	(0.003)	(0.001)
HH income	0.0002	0.001	−0.001	−0.001
	(0.002)	(0.001)	(0.002)	(0.001)
Parent trauma experience	−0.001	0.00003	0.0002	0.001
	(0.002)	(0.001)	(0.001)	(0.001)
Constant	0.074***	0.078***	0.043***	0.060***
	(0.016)	(0.009)	(0.014)	(0.009)
Observations	696	2,177	696	2,177
R^2	0.044	0.050	0.051	0.040
F Statistic	2.595***	8.790***	3.046***	6.893***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: Ability, Budget and household characteristics

	Dependent variable:							
	Ability Low		Ability High		Budget Low		Budget High	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
R3	0.004	−0.005	−0.019	−0.048***	−0.070***	−0.062***	−0.020*	−0.034***
	(0.012)	(0.006)	(0.013)	(0.007)	(0.019)	(0.010)	(0.011)	(0.006)
Machine annotated		0.002		0.005		−0.003		0.005
		(0.003)		(0.003)		(0.005)		(0.003)
Refugee	−0.015*	−0.018***	0.002	0.001	−0.036***	−0.030***	0.004	0.003
	(0.008)	(0.004)	(0.009)	(0.005)	(0.012)	(0.007)	(0.007)	(0.004)
Number of children	0.004*	0.004***	0.001	−0.001	0.007**	0.006***	−0.0002	−0.002*
	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.002)	(0.002)	(0.001)
Female HH head	0.006	0.004	−0.006	−0.009**	0.019*	0.004	0.0002	−0.005
	(0.007)	(0.004)	(0.008)	(0.004)	(0.011)	(0.006)	(0.006)	(0.003)
Age of HH head	0.0003	0.0001	−0.0005	−0.0004**	0.0001	0.0001	−0.0003	−0.0002
	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0005)	(0.0002)	(0.0003)	(0.0001)
Parent's years of education	−0.001*	−0.002***	0.003***	0.003***	−0.003***	−0.003***	0.002***	0.002***
	(0.001)	(0.0004)	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.0004)
Religiously educated parent	−0.016	−0.017**	0.002	0.007	0.006	−0.001	−0.0005	−0.002
	(0.014)	(0.008)	(0.015)	(0.009)	(0.021)	(0.012)	(0.012)	(0.007)
Female eldest child	−0.006	0.003	0.004	0.001	0.008	0.008*	−0.003	−0.002
	(0.006)	(0.003)	(0.006)	(0.003)	(0.009)	(0.005)	(0.005)	(0.003)
Age of eldest child	−0.0003	0.0001	0.0001	0.0003	0.0003	0.0003	−0.0001	0.0001
	(0.0004)	(0.0002)	(0.0005)	(0.0003)	(0.001)	(0.0004)	(0.0004)	(0.0002)
HH asset index	−0.005*	−0.004***	0.003	0.001	−0.005	−0.006***	0.003	0.001
	(0.002)	(0.001)	(0.003)	(0.001)	(0.004)	(0.002)	(0.002)	(0.001)
HH income	−0.002	−0.001	0.003*	0.001	−0.004	−0.002*	0.003**	0.001
	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
Parent trauma experience	−0.002**	−0.001*	0.001	0.001	0.001	−0.0002	0.0004	−0.0003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
Constant	0.041***	0.041***	0.052***	0.072***	0.106***	0.099***	0.038***	0.048***
	(0.013)	(0.007)	(0.014)	(0.008)	(0.020)	(0.011)	(0.011)	(0.006)
Observations	696	2,177	696	2,177	696	2,177	696	2,177
R^2	0.036	0.036	0.059	0.095	0.125	0.097	0.081	0.080
F Statistic	2.127**	6.205***	3.589***	17.400***	8.104***	17.841***	4.994***	14.381***

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

Table 9: Other Navigational Capacity variables and household characteristics

	Dependent variable:							
	Awareness Information Low		Awareness Information High		Reliance On God		Vague Non-specific	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
R3	−0.002	0.003	−0.038**	−0.024***	−0.023**	−0.014***	−0.037***	−0.040***
	(0.005)	(0.003)	(0.016)	(0.008)	(0.011)	(0.005)	(0.012)	(0.006)
Machine annotated		−0.003***		0.002		0.0002		0.003
		(0.001)		(0.004)		(0.003)		(0.003)
Refugee	0.001	−0.0002	0.0004	−0.008	−0.028***	−0.014***	−0.031***	−0.011***
	(0.004)	(0.002)	(0.010)	(0.006)	(0.007)	(0.004)	(0.008)	(0.004)
Number of children	−0.001	−0.001*	−0.003	−0.002*	−0.001	−0.001	−0.003	−0.001
	(0.001)	(0.0004)	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Female HH head	−0.003	0.001	0.010	0.003	0.010*	0.002	−0.008	−0.002
	(0.003)	(0.002)	(0.009)	(0.005)	(0.006)	(0.003)	(0.007)	(0.004)
Age of HH head	0.00000	0.0001*	−0.00001	0.00001	−0.0004	−0.0002	−0.0002	−0.0003*
	(0.0001)	(0.0001)	(0.0004)	(0.0002)	(0.0003)	(0.0001)	(0.0003)	(0.0002)
Parent's years of education	−0.001**	−0.0004**	0.003**	0.001**	−0.002**	−0.001***	−0.001	0.0003
	(0.0004)	(0.0002)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.001)	(0.0004)
Religiously educated parent	−0.002	−0.001	−0.003	−0.004	−0.011	−0.003	0.004	0.006
	(0.006)	(0.003)	(0.017)	(0.010)	(0.012)	(0.006)	(0.014)	(0.007)
Female eldest child	0.005*	0.004***	−0.001	−0.009**	0.007	0.003	0.003	−0.001
	(0.003)	(0.001)	(0.007)	(0.004)	(0.005)	(0.003)	(0.006)	(0.003)
Age of eldest child	0.0001	−0.00003	−0.00004	−0.0002	0.0001	0.0001	−0.001	−0.0001
	(0.0002)	(0.0001)	(0.001)	(0.0003)	(0.0004)	(0.0002)	(0.0004)	(0.0002)
HH asset index	−0.0005	0.0002	0.002	−0.002	−0.003	−0.0001	0.001	0.001
	(0.001)	(0.0005)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
HH income	0.0001	0.0002	0.003	0.002**	0.001	0.0001	−0.001	−0.001
	(0.001)	(0.0003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Parent trauma experience	0.0002	0.0001	0.001	0.001	0.0001	0.0003	−0.0002	0.0003
	(0.001)	(0.0003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.011*	0.005*	0.065***	0.073***	0.069***	0.052***	0.116***	0.084***
	(0.006)	(0.003)	(0.016)	(0.009)	(0.011)	(0.006)	(0.013)	(0.007)
Observations	696	2,177	696	2,177	696	2,177	696	2,177
R^2	0.023	0.016	0.069	0.038	0.069	0.028	0.166	0.121
F Statistic	1.347	2.664***	4.219***	6.603***	4.210***	4.705***	11.330***	22.981***

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

Table 10: Quant Education Ambition

	Dependent variable:					
	eld_edu_ambition					
	(1)	(2)	(3)	(4)	(5)	(6)
Refugee	−1.644***	−1.499***			−1.472***	−1.379***
	(0.216)	(0.117)			(0.203)	(0.111)
Female eldest child	−0.120	−0.313***			0.006	−0.189**
	(0.142)	(0.077)			(0.136)	(0.076)
Machine annotated				0.228**		0.099
				(0.089)		(0.076)
No Ambition			−10.972***	−9.305***	−4.469*	−5.138***
			(2.784)	(2.075)	(2.386)	(1.740)
Salaried Employment			2.663***	1.587***	1.793***	1.181***
			(0.637)	(0.361)	(0.573)	(0.314)
Vocational Training			−3.879*	−3.258*	−2.416	−2.335
			(1.978)	(1.709)	(1.657)	(1.429)
Entrepreneur			0.550	−0.686	−1.615*	−0.182
			(1.033)	(0.543)	(0.946)	(0.465)
Education Low			−4.184***	−5.363***	−1.429	−3.279***
			(1.564)	(1.040)	(1.347)	(0.915)
Education Neutral			−0.860*	−0.598**	−0.015	0.128
			(0.461)	(0.261)	(0.442)	(0.247)
Education High			3.636***	3.705***	2.606***	2.204***
			(0.810)	(0.495)	(0.747)	(0.441)
Education Religious			−3.264***	−3.396***	−1.385	−1.773***
			(0.962)	(0.591)	(0.850)	(0.522)
Marriage			−1.853***	−1.782***	−1.911***	−1.783***
			(0.709)	(0.391)	(0.661)	(0.359)
Migration			−0.045	−1.083	1.560	−0.330
			(1.257)	(0.817)	(1.390)	(0.760)
Constant	3.613***	3.987***	4.039***	4.171***	3.587***	3.917***
	(0.332)	(0.179)	(0.160)	(0.105)	(0.339)	(0.193)
Observations	392	1,184	426	1,267	392	1,184
R^2	0.411	0.389	0.286	0.206	0.515	0.466
F Statistic	24.089***	67.704***	16.610***	29.566***	18.672***	45.974***
Note:						*p<0.1; **p<0.05; ***p<0.01

Note: Coefficients on Number of children, Female HH head, Age of HH head, Parent's years of education, Religiously educated parent, Age of eldest child, HH asset index, HH income and Parent trauma experience are omitted to save space, full results are shown in Appendix C.

Table 11: Cost Benefit Scenarios

Objective	Budget	N_h	N_m	Price
Average F statistic	**$10,000**	**500**	**200**	$9,900
Ability Low Refugee effect	$10,000	500	200	$9,900
Secular Aspirations Female child effect	$10,000	100	600	$8,700
Average F statistic	**$15,000**	**500**	**600**	$14,700
Ability Low Refugee effect	$15,000	500	600	$14,700
Secular Aspirations Female child effect	$15,000	200	1,000	$15,000
Average F statistic	**$20,000**	**500**	**1,000**	$19,500
Ability Low Refugee effect	$20,000	300	1,200	$18,900
Secular Aspirations Female child effect	$20,000	300	1,200	$19,500

Table 12: 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 13: 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	Salaried Employment	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	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 14: 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, cluelessness.	• Question: What kind of doctor would you be happy with? Answer: He could be a popular doctor.

Table 15: 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

Table 16: Resolving disagreement

Code**Code**	**Description**	**AH**	**AK**	**MB**
Salaried Employment	Coded when secular 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

Table 17: Statistical methods for text vectorization

Method Name	Description	Hyperparameters (Options)	Hyperparameters (Used)
TfidfVectorizer	TfidfVectorizer is a method for converting text into numerical representations, specifically term frequency-inverse document frequency (TF-IDF) vectors. It counts the frequency of words in a document and down-weights the importance of commonly used words. This can be useful for text classification tasks, as it allows the classifier to focus on the words that are most distinctive to a particular document.	• **ngram_range**: The range of n-grams to consider when creating the vocabulary. • **min_df**: The minimum number of documents a word must be in to be included in the vocabulary. • **max_df**: The maximum number of documents a word can be in to be included in the vocabulary. • **max_features**: The maximum number of words to keep in the vocabulary, based on word frequency. • **use_idf**: A boolean flag indicating whether to use the inverse-document-frequency weighting. • **norm**: The type of normalization to apply to the vectors. • **smooth_idf**: A boolean flag indicating whether to smooth the idf values. • **sublinear_tf**: A boolean flag indicating whether to apply sublinear scaling to the term frequency.	• **max_features**: The maximum number of words to keep in the vocabulary based on word frequency. **[1000, 10000]** • **ngram_range**: The lower and upper boundary of the range of n-values for different word n-grams to be extracted. **{ (1,1), (1,2), (1,3) }**
CountVectorizer	CountVectorizer is a method for converting text into numerical representations, specifically a sparse matrix of word counts. It counts the frequency of words in a document and does not down-weight the importance of commonly used words. This can be useful for text classification tasks, as it allows the classifier to consider all words equally, rather than down-weighting the importance of commonly used words.	• **ngram_range**: The range of n-grams to consider when creating the vocabulary. • **min_df**: The minimum number of documents a word must be in to be included in the vocabulary. • **max_df**: The maximum number of documents a word can be in to be included in the vocabulary. • **max_features**: The maximum number of words to keep in the vocabulary, based on word frequency. • **binary**: A boolean flag indicating whether to create binary vectors, with 0/1 values indicating the presence/absence of a word in a document.	• **max_features**: The maximum number of words to keep in the vocabulary, based on word frequency. **[1000, 10000]** • **ngram_range**: The lower and upper boundary of the range of n-values for different word n-grams to be extracted. **{ (1,1), (1,2), (1,3) }** • **binary**: Whether to use binary or frequency counts. **{True, False}**

Table 18: Pre-trained embeddings for text vectorization

Model Name	**Dimensions**	**Description**
all-mpnet-base-v2	768	This a pre-trained language understanding model that combines the advantages of masked language modeling (MLM) and permuted language modeling (PLM) to address the limitations of both methods. It leverages the dependency among predicted tokens through PLM and takes auxiliary position information as input to make the model see a full sentence, reducing the position discrepancy between pre-training and fine-tuning. This model was pre-trained on a large-scale dataset and generates a vector of 768 dimensions.
all-roberta-large-v1	1024	This is a pre-trained language understanding model with a vector representation of 1024 dimensions. It was developed as an improvement upon the BERT model and was trained using the masked language modeling (MLM) objective. It has achieved strong performance on natural language processing tasks and can be fine-tuned on labeled datasets for specific tasks such as classification or language translation.
average_word_embeddings _glove.6B.300d	300	This is a method for converting text into numerical representations, specifically word embeddings. It uses a pre-trained GloVe model to generate 300-dimensional vector representations for each word in a document, and then averages these vectors to create a single representation for the entire document. This can be useful for text classification tasks, as it allows the classifier to consider the semantic relationships between words, rather than just their frequencies.
distiluse-base-multilingual-cased-v2	512	This is a pre-trained language understanding model that maps text into a 512-dimensional vector representation. It is a smaller and faster version of the popular transformer model, BERT, and has been trained on a large multilingual dataset, allowing it to process text in multiple languages. It has also been cased, meaning it can distinguish between upper and lower case letters. This model is useful for natural language processing tasks such as language translation and text classification, and can be fine-tuned on labeled datasets for specific tasks.

Table 19: Classifier Options I

Method	Description	Hyperparameters (Options)	Hyperparameters (Used)
LogisticRegression	This is a linear classifier that uses a logistic function to predict the probability of a sample belonging to a particular class. It is commonly used for binary classification tasks, but can also be used for multi-class classification by implementing a one-versus-rest approach.	• C: The inverse of the regularization strength, with higher values indicating less regularization. • penalty: The type of regularization to use, either L1 or L2. • fit_intercept: A boolean flag indicating whether to fit an intercept term. • tol: The tolerance for stopping criteria. • intercept_scaling: The scaling of the intercept term, if it is being fitted. • class_weight: The class weights to use for unbalanced classes. • max_iter: The maximum number of iterations for the optimization algorithm.	• penalty: The type of regularization to use: L1 or L2. • C: Inverse of regularization strength. [0.00002, 10000]
SGDClassifier	This is a linear classifier that uses stochastic gradient descent to learn the parameters of the model. The modified huber loss function is a smooth approximation of the hinge loss, which is commonly used for linear classification tasks.	• loss: The loss function to use, with options such as “hinge”, “log”, “modified_huber”, “squared_hinge”, and “perceptron”. • penalty: The type of regularization to use, with options such as L1, L2, “elasticnet”, and “none”. • alpha: The regularization strength, with higher values indicating stronger regularization. • l1_ratio: The proportion of L1 regularization to use in the elasticnet penalty. • tol: The tolerance for the stopping criteria. • learning_rate: The learning rate for the optimization algorithm, with options such as “constant”, “optimal”, and “invscaling”. • eta0: The initial learning rate for the “constant” and “invscaling” learning rate schedules. • power_t: The exponent for the “invscaling” learning rate schedule.	• loss: The loss function to use. (“modified_huber”) • penalty: The type of regularization to use: L1 or L2. • learning_rate: The learning rate schedule to use. (“optimal”) • alpha: The constant that multiplies the regularization term. [0.00002, 1000]

Table 20: Classifier Options II

Method	Description	Hyperparameters (Options)	Hyperparameters (Used)
RandomForestClassifier	This is an ensemble classifier that uses multiple decision trees to make predictions. It randomly selects a subset of features to consider at each split in the tree, which helps to reduce overfitting and improve the generalization of the model.	• n_estimators: The number of decision trees in the forest. • criterion: The function to measure the quality of a split, with options such as "gini" and "entropy". • max_depth: The maximum depth of the decision tree. • min_samples_split: The minimum number of samples required to split an internal node. • min_samples_leaf: The minimum number of samples required to be at a leaf node. • min_weight_fraction_leaf: The minimum weighted fraction of the sum total of weights required to be at a leaf node. • max_features: The number of features to consider when looking for the best split. • max_leaf_nodes: The maximum number of leaf nodes in the tree. • min_impurity_decrease: The minimum decrease in impurity required to split the node. • bootstrap: A boolean flag indicating whether to use bootstrap samples when building the trees. • oob_score: A boolean flag indicating whether to use out-of-bag samples to estimate the generalization error.	• n_estimators: The number of trees in the forest. [100, 1000] • max_depth: The maximum depth of the tree. [10, 100]
DecisionTreeClassifier	This is a classifier that uses a tree structure to make decisions based on the features of a sample. At each node in the tree, the classifier considers a single feature and splits the data based on the value of that feature. The final decision is made based on the path taken through the tree.	• criterion: The function to measure the quality of a split, with options such as "gini" and "entropy". • splitter: The strategy to use when searching for a split, with options such as "best" and "random". • max_depth: The maximum depth of the tree. • min_samples_split: The minimum number of samples required to split an internal node. • min_samples_leaf: The minimum number of samples required to be at a leaf node. • min_weight_fraction_leaf: The minimum weighted fraction of the sum total of weights required to be at a leaf node. • max_features: The number of features to consider when looking for the best split. • max_leaf_nodes: The maximum number of leaf nodes in the tree. • min_impurity_decrease: The minimum decrease in impurity required to split the node.	• max_depth: The maximum depth of the tree. [5, 100] • min_impurity_decrease: A node will be split if this split induces a decrease of the impurity greater than or equal to this value. [0.00002,10000]

Table 21: Classifier Options III

Method	Description	Hyperparameters (Options)	Hyperparameters (Used)
MLPClassifier	This is a classifier that uses a neural network with multiple layers to make predictions. It is commonly used for classification tasks and can handle both continuous and categorical data. The number of layers and the number of units in each layer can be adjusted to fit the complexity of the task.	• hidden_layer_sizes: The number of neurons in each hidden layer. • activation: The activation function to use, with options such as "identity", "logistic", "tanh", and "relu". • solver: The algorithm to use for optimization, with options such as "lbfgs", "sgd", and "adam". • alpha: The regularization strength, with higher values indicating stronger regularization. • batch_size: The number of samples to use in each iteration of the optimization algorithm. • learning_rate: The learning rate for the optimization algorithm, with options such as "constant", "invscaling", and "adaptive". • learning_rate_init: The initial learning rate for the "constant" and "invscaling" learning rate schedules. • power_t: The exponent for the "invscaling" learning rate schedule. • max_iter: The maximum number of iterations to run the optimization algorithm. • shuffle: A boolean flag indicating whether to shuffle the training data before each epoch. • tol: The tolerance for the stopping criteria. • warm_start: A boolean flag indicating whether to reuse the solution of the previous call to fit. • momentum: The momentum for the optimization algorithm. • nesterovs_momentum: A boolean flag indicating whether to use Nesterov's momentum. • early_stopping: A boolean flag indicating whether to use early stopping to terminate the optimization early. • validation_fraction: The fraction of the training data to use as validation data for early stopping. • beta_1: The beta 1 parameter for the Adam optimization algorithm.	• hidden_layer_sizes: The ith element represents the number of neurons in the ith hidden layer. [(100,), (100, 100), (100, 100, 100)] • activation: Activation function for the hidden layer. ("tanh", "relu") • alpha: L2 penalty (regularization term) parameter. [0.01, 1]

Table 22: Classifier Options III

Method	Description	Hyperparameters (Options)	Hyperparameters (Used)
KNeighborsClassifier	This is a non-parametric classifier that uses the K nearest neighbors of a sample to make a prediction. It is commonly used for classification tasks and can handle both continuous and categorical data. The number of neighbors to consider (K) is a hyperparameter that can be adjusted to fit the complexity of the task.	• n_neighbors: The number of neighbors to use when making a prediction. • weights: The weight function to use when making a prediction, with options such as "uniform" and "distance". • algorithm: The algorithm to use for finding the nearest neighbors, with options such as "brute" and "kd_tree". • leaf_size: The number of points at which to switch to a brute force search for the nearest neighbors. • p: The power parameter for the Minkowski distance metric. • metric: The distance metric to use, with options such as "euclidean", "manhattan", and "minkowski". • metric_params: Additional parameters for the distance metric.	• n_neighbors: Number of neighbors to use by default for kneighbors queries. [10,10000] • weights: weight function used in prediction. ("uniform", "distance")
SVC	This is a classifier that uses a support vector machine (SVM) to find the optimal hyperplane to separate the different classes. It is commonly used for classification tasks and can handle both continuous and categorical data. The kernel function used to project the data into a higher dimensional space can be adjusted to fit the complexity of the task.	• C: The regularization strength, with higher values indicating stronger regularization. • kernel: The kernel to use for the decision function, with options such as "linear", "poly", "rbf", "sigmoid", and "precomputed". • degree: The degree of the polynomial kernel. • gamma: The kernel coefficient for the rbf, poly, and sigmoid kernels. • coef0: The independent term in the polynomial and sigmoid kernels. • shrinking: A boolean flag indicating whether to use the shrinking heuristic. • probability: A boolean flag indicating whether to enable probability estimates. • tol: The tolerance for the stopping criteria. • class_weight: The class weights to use for unbalanced classes. • verbose: The level of verbosity in the output. • decision_function_shape: The shape of the decision function, with options such as "ovo" and "ovr".	• C: Penalty parameter C of the error term. [0.00001, -00]

Table 23: Annotations with evidence of bias

	Dependent variable:		
	No Ambition errors	Education Neutral errors	Awareness Information Low errors
	(1)	(2)	(3)
R3	−0.002	0.056***	−0.001
	(0.006)	(0.018)	(0.006)
Refugee	0.012***	0.025**	0.009**
	(0.004)	(0.012)	(0.004)
Number of Children	−0.001	−0.003	−0.001
	(0.001)	(0.003)	(0.001)
Female HH head	−0.006*	0.014	−0.004
	(0.003)	(0.011)	(0.003)
Age of HH head	0.0002	0.0003	−0.0001
	(0.0001)	(0.0005)	(0.0001)
Parent's years of education	−0.0002	−0.0002	−0.001***
	(0.0004)	(0.001)	(0.0004)
Religiously educated parent	−0.005	−0.008	−0.011*
	(0.006)	(0.020)	(0.006)
Female eldest child	0.001	−0.005	0.002
	(0.003)	(0.008)	(0.003)
Age of eldest child	0.0001	−0.001	0.0001
	(0.0002)	(0.001)	(0.0002)
HH asset index	−0.001	0.005	0.001
	(0.001)	(0.003)	(0.001)
HH income	−0.0001	−0.001	0.0004
	(0.001)	(0.003)	(0.001)
Parent trauma experience	−0.0001	0.001	0.0002
	(0.001)	(0.002)	(0.001)
Constant	−0.009	−0.050***	0.005
	(0.006)	(0.019)	(0.006)
Observations	696	696	696
R^2	0.057	0.043	0.038
F Statistic	3.421***	2.585***	2.251***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 24: Quant Ambition: full results with coefficients for all quant variables

	Dependent variable:					
	eld_edu_ambition					
	(1)	(2)	(3)	(4)	(5)	(6)
Refugee	−1.644***	−1.499***			−1.472***	−1.379***
	(0.216)	(0.117)			(0.203)	(0.111)
Number of children	0.119**	0.028			0.113**	0.022
	(0.055)	(0.029)			(0.051)	(0.028)
Female HH head	0.134	0.106			0.134	0.059
	(0.198)	(0.101)			(0.186)	(0.095)
Age of HH head	0.016**	0.012***			0.012*	0.009**
	(0.007)	(0.004)			(0.007)	(0.004)
Parent's years of education	0.064***	0.074***			0.029	0.056***
	(0.020)	(0.011)			(0.019)	(0.011)
Religiously educated parent	0.271	0.676***			0.282	0.768***
	(0.364)	(0.201)			(0.337)	(0.190)
Female eldest child	−0.120	−0.313***			0.006	−0.189**
	(0.142)	(0.077)			(0.136)	(0.076)
Age of eldest child	0.004	0.003			0.003	0.005
	(0.006)	(0.003)			(0.006)	(0.003)
HH asset index	0.088	0.075**			0.070	0.058**
	(0.058)	(0.030)			(0.053)	(0.028)
HH income	−0.003	0.015			0.010	0.013
	(0.037)	(0.017)			(0.035)	(0.016)
Parent trauma experience	0.007	0.030*			0.014	0.029*
	(0.031)	(0.017)			(0.029)	(0.017)
Machine annotated				0.228**		0.099
				(0.089)		(0.076)
No Ambition			−10.972***	−9.305***	−4.469*	−5.138***
			(2.784)	(2.075)	(2.386)	(1.740)
Salaried Employment			2.663***	1.587***	1.793***	1.181***
			(0.637)	(0.361)	(0.573)	(0.314)
Vocational Training			−3.879*	−3.258*	−2.416	−2.335
			(1.978)	(1.709)	(1.657)	(1.429)
Entrepreneur			0.550	−0.686	−1.615*	−0.182
			(1.033)	(0.543)	(0.946)	(0.465)
Education Low			−4.184***	−5.363***	−1.429	−3.279***
			(1.564)	(1.040)	(1.347)	(0.915)
Education Neutral			−0.860*	−0.598**	−0.015	0.128
			(0.461)	(0.261)	(0.442)	(0.247)
Education High			3.636***	3.705***	2.606***	2.204***
			(0.810)	(0.495)	(0.747)	(0.441)
Education Religious			−3.264***	−3.396***	−1.385	−1.773***
			(0.962)	(0.591)	(0.850)	(0.522)
Marriage			−1.853***	−1.782***	−1.911***	−1.783***
			(0.709)	(0.391)	(0.661)	(0.359)
Migration			−0.045	−1.083	1.560	−0.330
			(1.257)	(0.817)	(1.390)	(0.760)
Constant	3.613***	3.987***	4.039***	4.171***	3.587***	3.917***
	(0.332)	(0.179)	(0.160)	(0.105)	(0.339)	(0.193)
Observations	392	1,184	426	1,267	392	1,184
R^2	0.411	0.389	0.286	0.206	0.515	0.466
F Statistic	24.089***	67.704***	16.610***	29.566***	18.672***	45.974***
Note:						*p<0.1; **p<0.05; ***p<0.01

List of Figures

Figure 1: Coding tree

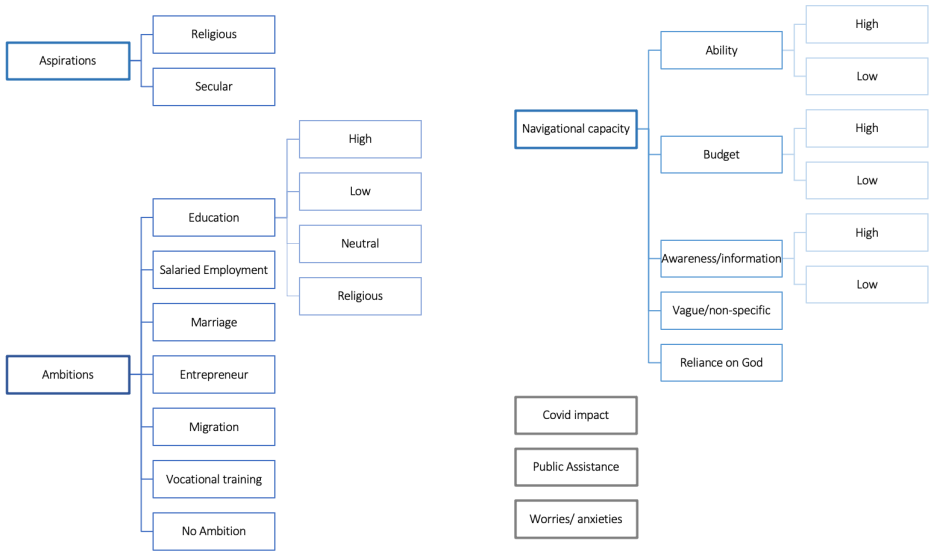


Figure 2: Examples of qualitative codes

(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).."

Figure 3: Methodology

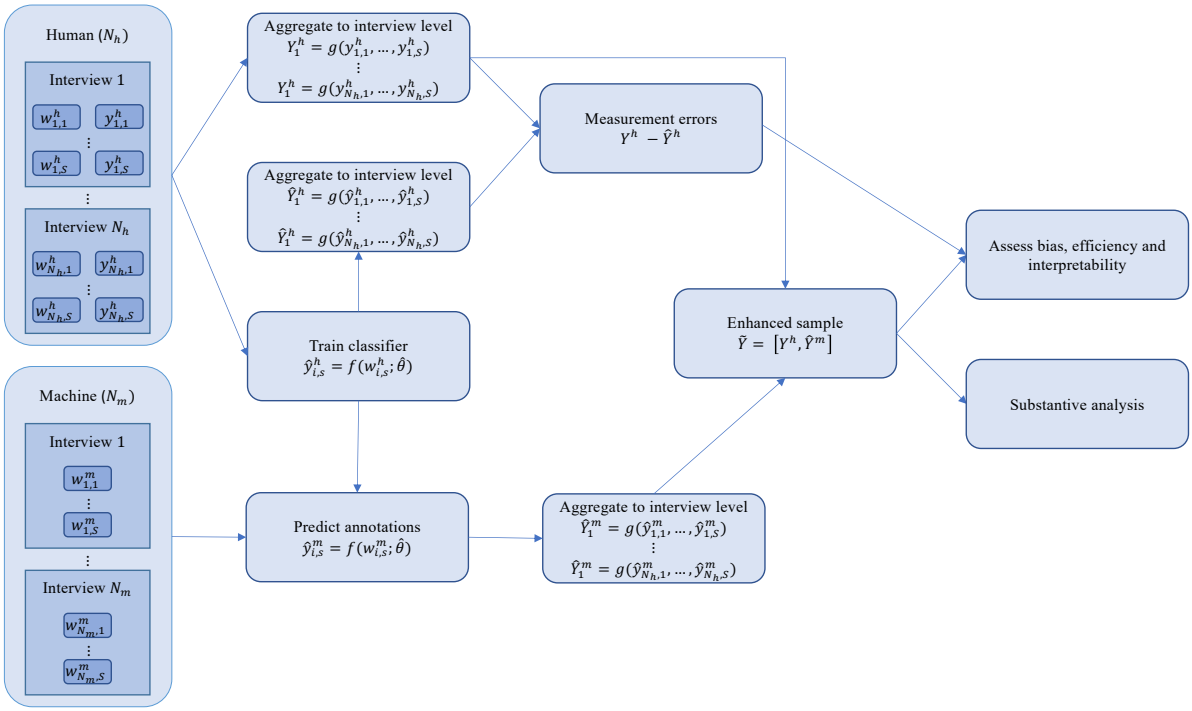
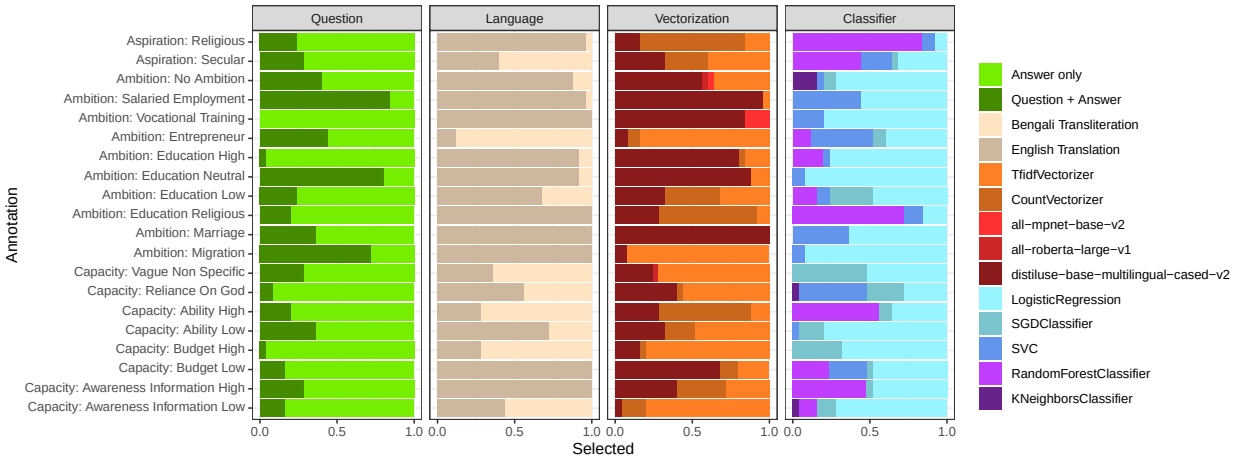
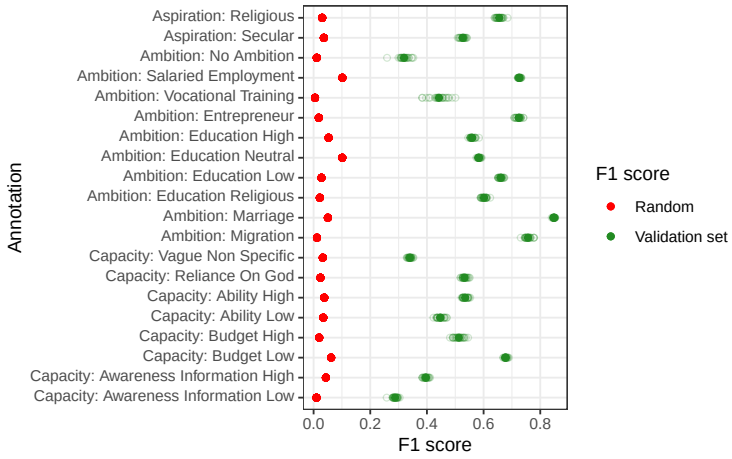


Figure 4: Choices of text representation and classifier



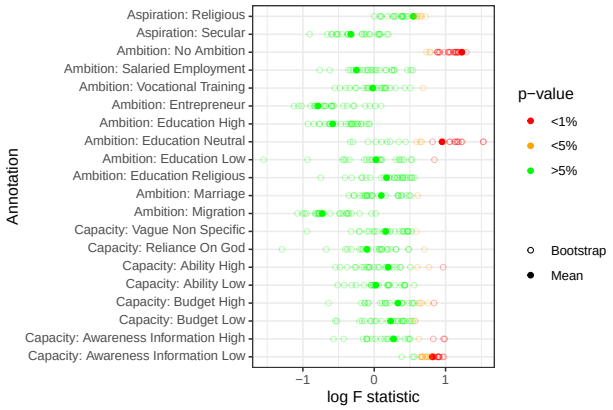
Note: This Figure shows the selected text representation and classifier for each annotation across 25 bootstraps. The first panel shows the proportion of runs in which the Question is included in the text representation. The second panel shows whether the chosen text representation was based on Bengali transliterated into Latin characters, or a machine translation into English. The third panel shows the selected vectorizer, which is applied to convert the text into numeric vectors. Finally, the fourth panel shows the selected classifier.

Figure 5: Validation set performance



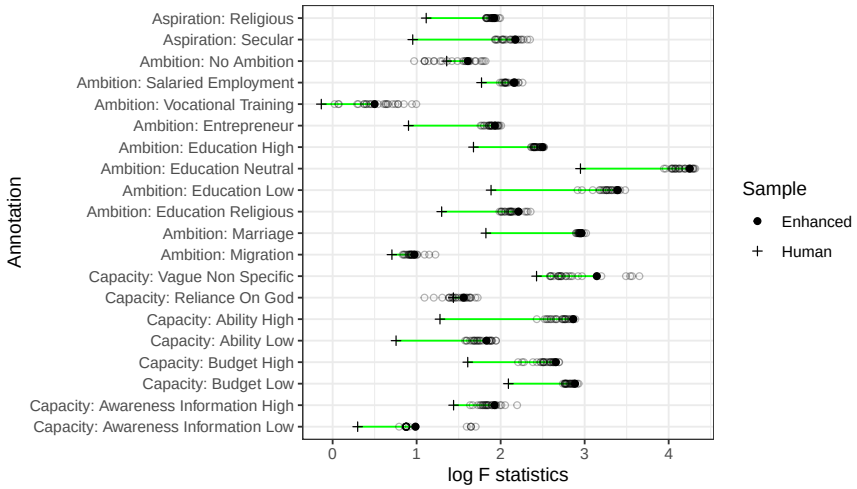
Note: This Figure shows validation set performance of the selected model for each annotation and each bootstrap run, as measured by the F1 score. The sparsity of the annotation across QA pairs is shown in red as a reference point: this would be the expected F1 score if predictions were drawn randomly based on the overall proportion of positives.

Figure 6: Bias test for each annotation



Note: This Figure shows the log F statistic for the regression of the validation set errors on household characteristics, for each annotation. The color of each point indicates the significance level of the F statistic. The hollow circles represent the statistic for each bootstrap and the solid circle represents the statistic for an enhanced sample based on the mean prediction across each bootstrap.

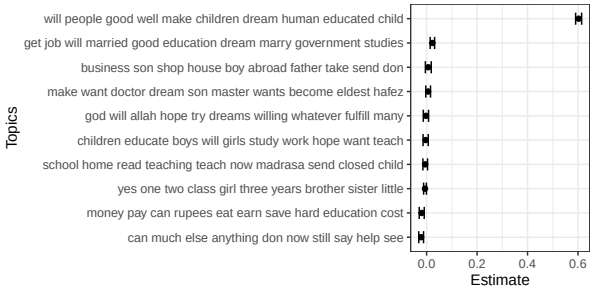
Figure 7: Interpretability test



Note: This Figure shows the log F statistic for the regression of each annotation on household characteristics in the enhanced and human samples. The hollow circles represent the statistic for each bootstrap and the solid circle represents the statistic for an enhanced sample based on the mean prediction across each bootstrap.

Figure 8: Example of supervised LDA topics

(a) Aspirations:Secular



(b) Aspirations:Religious

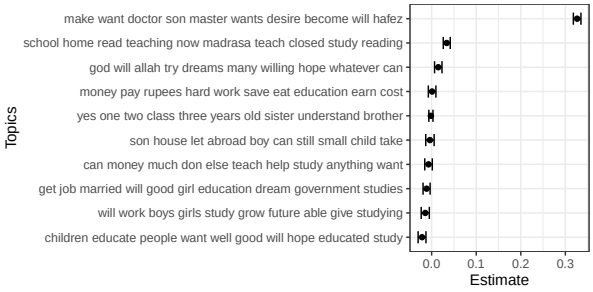
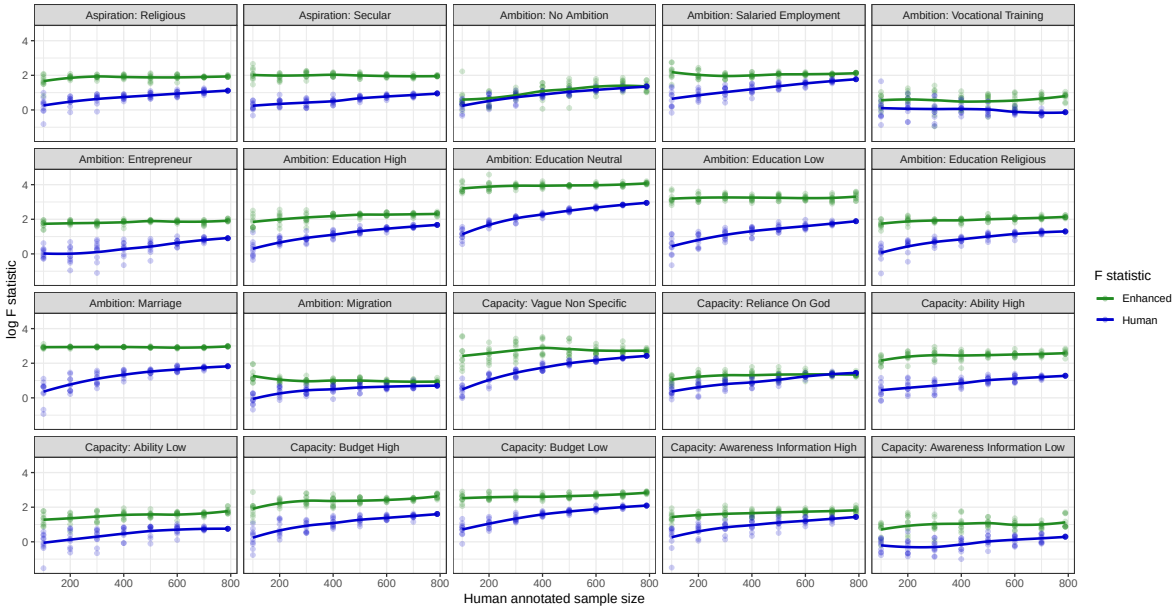


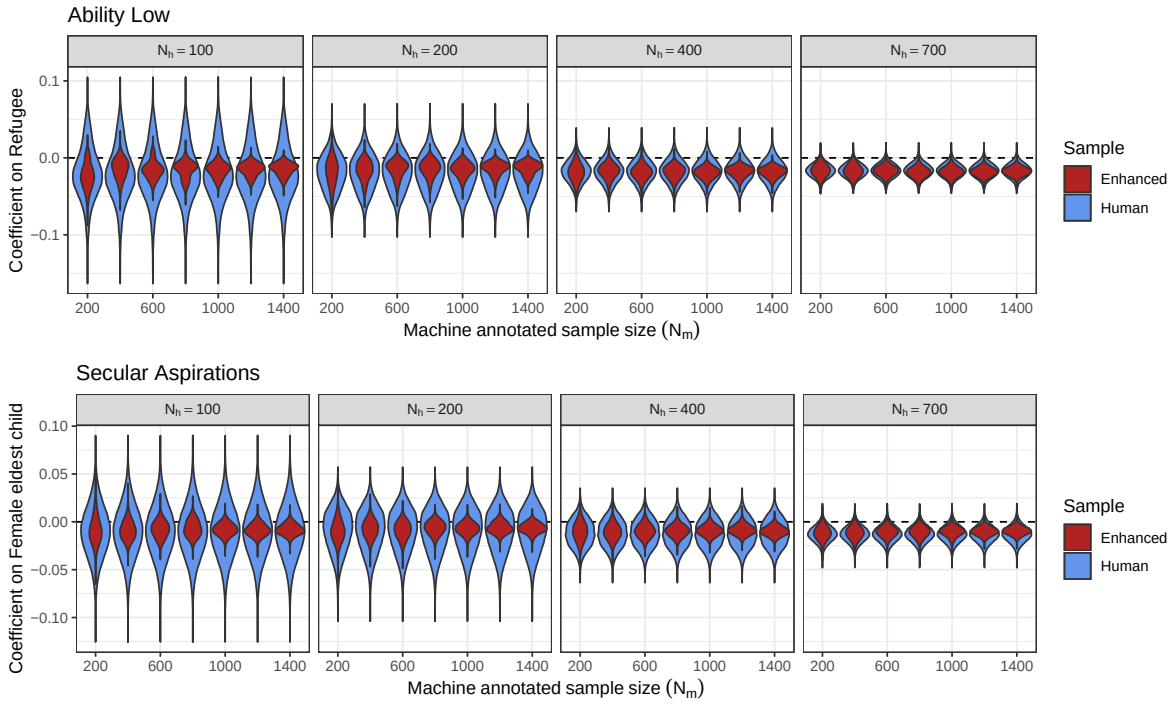
Figure 9: Correlations between annotations in enhanced sample

Figure 10: F-statistic test for interpretability increases with N_h (holding N fixed)



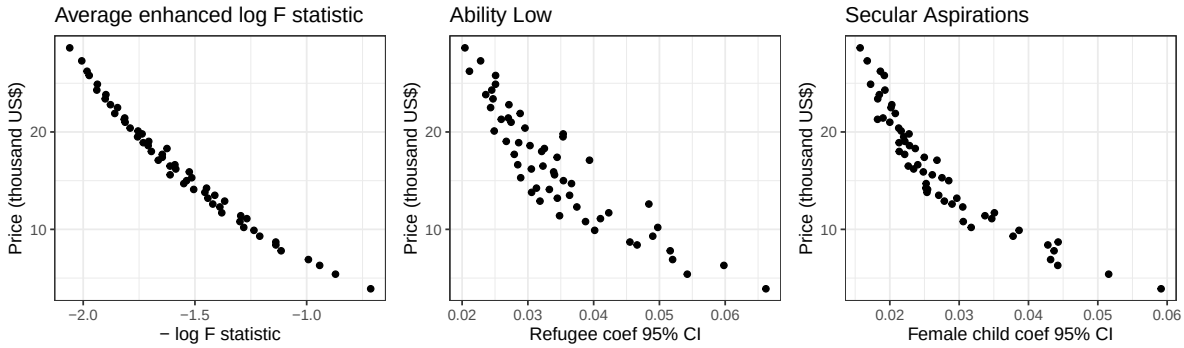
Note: This Figure shows F statistics of for each annotation of a regression on household characteristics in the human (in blue) and enhanced (in green) samples as existing interviews are annotated. The total sample size in the enhanced sample is thus constant, but interviews are moved from the machine annotated set to the human annotated. Each point represents a bootstrap run and the lines show a local regression fit to these points.

Figure 11: Distribution of regression coefficients of interest with N_h and N_m



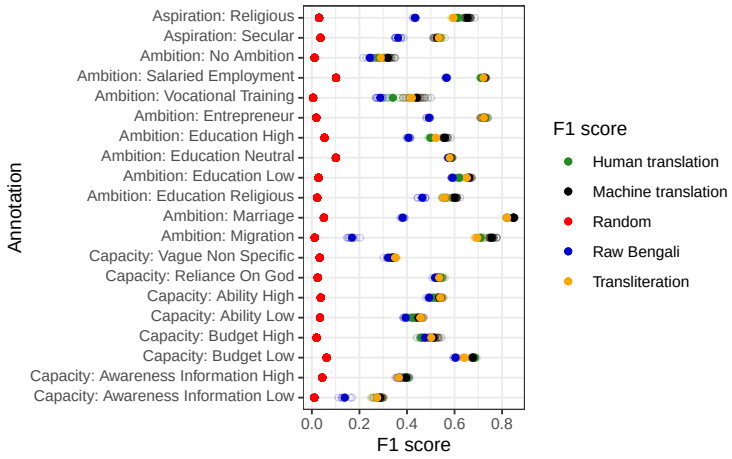
Note: This Figures shows how the distribution of coefficient estimates for two coefficients of interest change as the number of human annotated interviews (N_h) and the number of machine annotated interviews (N_m) are varied. The upper panels represent the coefficient on the refugee status variable in the regression for Ability Low, controlling for other household characteristics, i.e. from the first two columns of Table 8. The lower panels represent the coefficient on female eldest child variable in the regression for Secular Aspirations, i.e. from the first two columns of Table 7. In each case, the distribution of the coefficient estimated on the human annotated sample is shown in blue and on the enhanced sample is shown in red. Across panels, from left to right we show the effect of an increase in N_h, so within each panel the blue distribution is the same. As N_h increases we see that the coefficient estimated on the human sample becomes more precise. Within the panels, from left to right we show the effect of an increase in N_m. As N_m increase the coefficient estimated on the enhanced sample becomes more precise. Interestingly, the estimated coefficient in the enhanced sample for a large N_m does not vary much with N_h, suggesting that a large human annotated sample is not necessary to get value from the enhancement. The coefficient distributions are calculated through bootstrapping both which interviews are included in the training sample and the coefficient estimate itself.

Figure 12: Cost trade-offs



Note: This Figure plots each combination of N_h and N_m with the objective on the horizontal axes and the price on the vertical axes. Across all three panels, moving further to the south west indicates a cheaper combination and a better outcome. The first panel uses the average enhanced sample F statistic for a regression of annotation on household characteristics, across all annotations. The objective in the second and third panels respectively are the coefficient on refugee status for Ability Low and the coefficient on a female eldest child for Secular Aspirations.

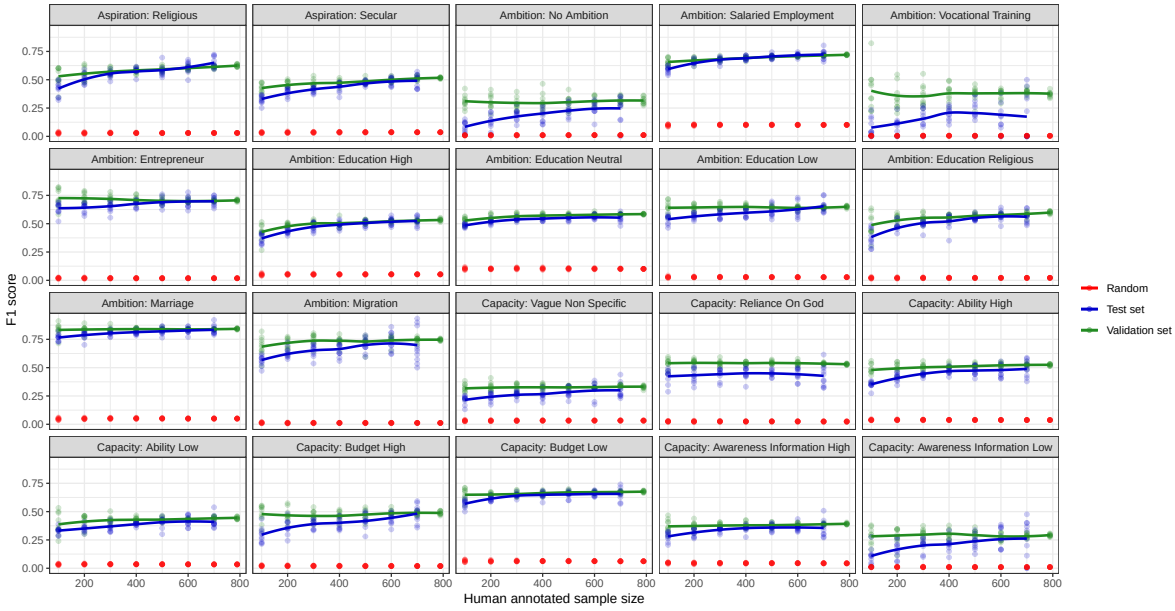
Figure 13: Validation set performance across different translation approaches



Note: Figure shows the validation set performance across different translation approaches. As in Figure 5, teh sparsity of each annotation is shown in red as a reference point. In each translation approach, we select over the possible vectorizers as described in Section 5.2. The average validation F1 scores across all annotations are 0.558 for Machine translation, 0.542 for Transliteration, 0.535 for Human translation and 0.420 for Raw Bengali.

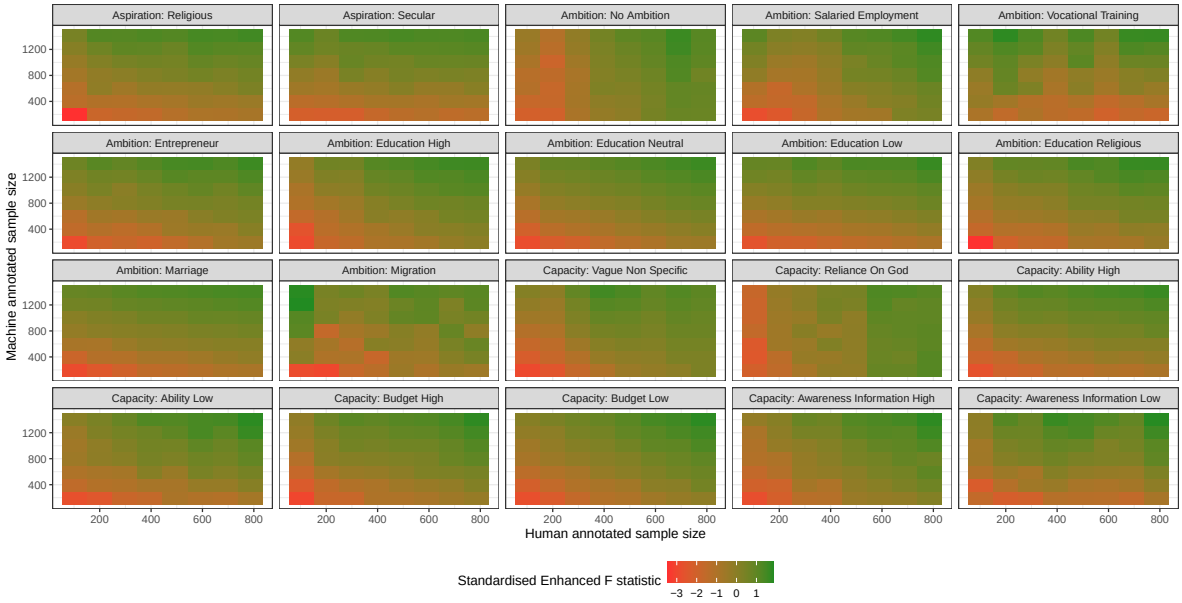
Figure 14: Correlations between annotations in human sample

Figure 15: Validation and test set performance for increasing N_h



Note: This Figure shows validation set performance (in green) and held-out test set performance (in blue) for each annotation as the size of the human annotated training set increases along the horizontal axes. Each point represents a bootstrap run and the lines show a local regression fit to these points. The sparsity of the annotation across QA pairs in the training set is shown in red as a reference point.

Figure 16: F-statistic test for interpretability increases with both N_h and N_m



Note: This Figure shows how interpretability of each annotation, as measured by the F statistic in a regression of the annotation on household characteristics in the enhanced sample, varies with the number of the human annotated interviews along the horizontal axis (N_h) and the number machine annotated interviews (N_m) along the vertical axis. The color of each cell corresponds to the mean F statistic in the enhanced sample across draws for that N_m and N_h. These F statistics are standardised so that they have zero mean and unit standard deviation within each annotation, ensuring a consistent color gradient for each annotation.