Qualitative Analysis with Large-N: A New Method with an Application to Aspirations in Bangladesh

Qualitative Analysis with Large-N

Julian Ashwin[1], Vijayendra Rao[2,*], Monica Biradavolu[3], Aditya Chhabra[2], Arshia Haque[2], Afsana Khan[2], and Nandini Krishnan[2]

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 labour intensive and very time consuming. This paper presents a method to simplify and shorten the coding process by extending a small sample of interpretative human-annotated interviews to a larger, representative, sample using supervised natural language processing. We extensively assess the robustness and reliability of this approach to show when and how it adds value, including an analysis of how many human-annotated documents are optimal for a budget constrained researcher. We apply this approach to analyse 2,200 open-ended interviews on parent's aspirations for children with

*__Correspondence address:__ vrao@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, the World Bank's "Half-Baked" seminar, the October 2022 meeting of the CIFAR Boundaries, Membership and Belonging Program, the LSE Inequalities Seminar, Erich Battistin and Ikechi Okorie for their useful comments. We are particularly grateful to the editor Barbara Petrongolo and four anonymous referees for valuable feedback on a previous version of this paper. Peer Nagy, Eleni Kalamara and Sudarshan Aittreya provided excellent research assistance for the project. The authors are grateful to the World Bank's Knowledge for Change Program, its Research Support Budget, and the World Bank-UNHCR Joint Data Center on Forced Displacement for financial support.

Rohingya refugees and their Bangladeshi hosts. We show 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. This approach allows us to identify several novel results including a new type of migrant selection; we show that Rohingya refugees are negatively selected on education but positively selected on navigational capacity.

Keywords: qualitative analysis; natural language processing; aspirations; bangladesh; rohingya

Classification: C63, Z10, O12, I25

1 Introduction

The qualitative analysis of open-ended interviews has vast potential in economics but has found limited use. We typically analyse quantitative data from structured survey questions because they are easier to administer to large representative samples of respondents, and easier to analyse using standard econometric methods. However, many questions of interest to economists may be better captured with open-ended 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 - in the case of this paper - aspirations.

This paper outlines a new method to analyse extended, conversational and open-ended interviews at scale with statistically representative samples by combining interpretative human coding (typically employed by sociologists) and machine learning. We extend a small sample of interpretative human-annotated interviews to a much large sample using supervised natural language processing. Using this method, we show that measuring aspirations with open-ended interviews extends the economic 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.

Structured quantitative questions do not work well for complex concepts like "moral aspirations" 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 nuanced are, therefore, arguably better studied with open-ended questions where the respondent is allowed the freedom to respond in a conversational style and a trained interviewer can probe and clarify an issue. 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 analysing them is hard using standard methods (Rao, 2022). Some economists have proposed using unsupervised NLP methods (e.g. Ferrario and Stantcheva (2022b), Stantcheva (2020)) and Large Language Models (LLMs) are increasingly being used in social science research to analyse and annotate text data (Gilardi *et al.*, 2023; Liu and Sun, 2023). Our paper proposes a different method that we believe is an improvement on unsupervised NLP for many applications and, in particular, for analysing open-ended interviews. It also provides a better alternative to Large Language Model (LLM) based qualitative analysis, which can introduce misleading biases in non-Western settings (Ashwin *et al.*, 2025).

Our method scales-up the "interpretive" approach to qualitative data analysis that is currently primarily the domain of "traditional" qualitative researchers in anthropology, sociology and related fields. In interpretive qualitative methods researchers take a more human-centred approach by mulling 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. This allows for a more nuanced, detailed and interpretable analysis of qualitative data but it also requires an expensive labour-intensive coding process typically done by trained social scientists. Traditional interpretative qualitative analysis is consequently associated with very small sample studies. This small-sample challenge that has been intrinsic to interpretive 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 non-representative samples 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).

The method outlined in this paper follows the logic of traditional interpretative qualitative analysis as closely as possible. Briefly, a sub-sample of the transcripts of open-ended interviews are coded by researchers trained in qualitative analysis who read the transcripts, decide on a "coding-tree" and then code the transcripts using qualitative analysis software 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 analysed 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 other social scientists. Our method allows social scientists to analyse 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. The key advantage of this larger sample size is that it can lead to smaller standard errors and give researchers the power to conduct more sophisticated and granular analyses. We see this as an organic extension of traditional, interpretative, human-coded qualitative analysis, but done at scale.

This method has several advantages over existing methods used for analysing text, which are discussed explicitly in Section 6. We stick as closely as possible to traditional qualitative analysis by inductively using the judgement of informed human coders to be scaled-up. Given that analysis is based on codes developed by a close reading of the text, the interpretation of our measures is clear, which is often not the case for unsupervised methods.[1] 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

[1]An explicit comparison to unsupervised methods is provided in Section 6.3.

the sample being analysed, and thus has the potential for nuanced, context-specific analysis. We also show that our method has advantages over LLMs which can result in biased results when analysing data from data (such as ours) from interactive open-ended interviews in non-Western populations.

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

Aspirations are a useful topic to test the method's potential because it allows us to measure aspects of the subject that have not been empirically analysed to date because they are difficult to capture in structured questionnaires. In a recent literature review, Fruttero *et al.* (2025) point out that dimensions of aspirations that go beyond material standards of living are not easily translated into quantifiable measures, saying that "measures often focus on a limited set of quantifiable measures (e.g. years of education, earnings, etc.) which are certainly central but may only partially reflect people's aspirations." Thus, despite theoretical advances in economics that take a broader approach to the idea of aspirations (Genicot and Ray, 2020) the empirical literature focuses on what the philosopher Agnes Callard (2018) has called "ambition" - specific material goals that parents may have for their children such as a level of education, potential earnings or occupation. Open-ended interviews allow us to expand this, to explore its normative, moral and spiritual dimensions - what Callard calls "aspiration" - such as being a "good person", or being religiously inclined. Our approach also allows us to measure what the anthropologist Arjun Appadurai (2004) has called "navigational capacity", which he conceives of as a culturally-shaped and unequally-distributed capacity towards achieving aspirational goals. This potentially important idea has been relatively ignored by the economics literature arguably because it is difficult to measure. This paper uses the analysis of open-ended interviews to define, measure and investigate correlates of these subtle concepts, in particular to understand their gender dimensions, how they differ between refugees and hosts, and the degree to which they are intergenerationally transmitted and thus contribute to the perpetuation of inequality. We demonstrate that these concepts have distinctly different determinants, and expand our understanding not just of aspirations but other topics such as migrant selection and inter-generational mobility.

Specifically, we find that the open-ended interviews provide responses on ambition that are in line with the previous literature. However, on the normative and spiritual dimensions of aspirations we find that parents are less likely to express both "secular" and "religious" aspirations for their female children, and more educated parents' aspirations are more likely to be secular than religious. This expands our understanding of gender discrimination in aspirations, and adds insights into the intergenerational transmission of norms and values.

Further, while Rohingya refugees are much less educated than their Bangladeshi hosts they nevertheless display higher navigational capacity. Refugees are less likely to display low navigational capacity, across several dimensions, than local Bangladeshis. This is consistent with a selection effect because of the considerable hardships they have faced in fleeing conflict and finding their way to the camps. This introduces a potentially significant new dimension to the literature on migrant selection which has previously focused on education and earnings. Even if refugees are less educated they may be more "capable" than their hosts - which could matter for policy in conflict situations. Our key substantive results are discussed in the context of the relevant economic literature in Section 5.

Related literature. The difficulties with using qualitative methods at scale on representative samples have led to their largely being neglected in modern economics.[2] 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). Much of this work has focused on "observational" text data like news articles, political speeches and patent filings, but increasingly economists are also making use of unsupervised NLP methods to allow analysis of open-ended survey responses (Stantcheva, 2020; Houde and Wekhof, 2021; Filippini *et al.*, 2021). These paper mostly focus on analyse open-text fields in online surveys, which are more structured and less complex than our transcripts of interactive interviews, in the Rohingya language and two different dialects of Bengali, that we work with. Economists are recognising the need to directly measuring interview subjects' thoughts and beliefs in a way that allows them to express themselves freely, as surveyed in Haaland *et al.* (2024), and we see our paper as contributing towards this end.

While we focus on the analysis of interviews conducted with human interviewers, there is also increasing interest in using artificial intelligence to conduct qualitative interviews (Chopra and Haaland, 2023; Geiecke and Jaravel, 2024). This is certainly an interesting avenue to explore, and could indeed be combined with our approach to analysing those interviews, but we would emphasise that the bias issues we find when using LLMs for analysis, as discussed in 6.2 and Ashwin *et al.* (2025), may also be in issue when conducting interviews, particularly in the context of topics and populations that are not well represented in LLMs' training data.

[2]There are notable exceptions, such as the widely used monetary policy shock series developed by Romer and Romer (2004) that relies on detailed readings of central bank minutes.

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.[3] 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 (Baker *et al.*, 2016; Shapiro *et al.*, 2020), but our focus is on using the manual classifications to construct a measure.[4] Michalopoulos and Xue (2021) use an archive of manually coded motifs in folklore introduced by Berezkin (2015) and then use NLP to classify these motifs into different concepts. A recent paper by Jayachandran *et al.* (2021) is similar to ours in spirit, as they use a subset of manually coded documents in order to identify which quantitative survey questions best capture women's agency. Although their approach is methodologically quite different the aim is similar: to 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 - a potential use case of the methodology we discuss here.[5]

There is a related literature outside economics on training supervised models on human annotations. However, this typically focuses on either maximizing predictive performance or assisting an ongoing coding process, rather than whether and how such methods can assist

[3]In Section 7 we extensively explore the effect of varying the size of this annotated set, and find that, in our application at least, annotating even a relatively small number of interviews and scaling these up gives most of the benefits of our method.
[4]There are also examples of manually annotating large samples, e.g. Andre *et al.* (2024) use open-ended survey responses to measure narratives about the macroeconomy, but rely on research assistants to annotate their entire sample.
[5]Using basic human-coding to create a training set has been used by a few others to analyse large corpora of secondary text data (e.g. Bonikowski and DiMaggio (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.

substantive economic analysis.[6] Yordanova *et al.* (2019) provide a summary of the literature that focuses on predictive performance which generally aims to show that a particular modelling approach yields superior predictive performance in these tasks. In contrast to this literature, 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. 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 of interest 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 annotations, and then use these to crowd-source 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 crowd-sourced 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. Our approach also contrasts with the growing literature that assesses the value of Large Language Models for annotating text documents, which we explicitly address in Section 6.2.

The paper proceeds next by giving a general presentation of the Method in Section 2. We then introduce our example application relating to aspirations of Rohingya refugees and their

[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.

Bangladeshi hosts, describing our data and qualitative coding process in Section 3. Section 4 then discusses our results from a methodological perspective, including a range of tests that assess the value of the enhanced sample. Section 5 then shows and discusses the key substantive results of our application to aspirations. Section 6 then compares our method to alternative approaches including traditional quantitative survey questions and other tools from the NLP literature. Section 7 then describes a series of experiments in which we assess how the number of human and machine annotated documents affect results. Finally, Section 8 concludes and makes suggestions for further work.

2 General Presentation of the Method

In this Section, we describe our approach. We give a brief introduction to qualitative analysis, before describing our strategy in general terms. We then discuss details and options for the supervised models, text representations and training method we use. Finally, we discuss assessing the value the enhanced sample with tests for out-of-sample performance, bias and efficiency.

2.1 Qualitative Analysis

The approach to data collection and analysis emphasized in this paper is bottom-up and reflexive Rao (2022) aiming to give respondents as much freedom as possible in shaping their answers, unlike the more rigid structure typical of standard quantitative surveys. Interviewers are trained to begin with open-ended prompts, such as: “What are your hopes and dreams for your children?” Respondents may reply with a few words, short phrases, or extended narratives. Interviewers are instructed to listen attentively and pose follow-up questions that encourage elaboration or clarification without leading or steering the conversation. The goal is a natural, organically unfolding dialogue.

In disciplines like anthropology and sociology, where this style of data collection is common, the interview transcripts are subsequently analysed through a process of summarization and coding. This method—referred to as flexible coding Detering and Waters (2018) and that enables researchers to use a predefined set of codes (often derived from theory) while also allowing new codes to emerge inductively from close textual analysis. In this way, researchers construct a hierarchical coding framework or coding tree, using both deductive and inductive reasoning.

Ideally, coding is conducted collaboratively in a team to mitigate the influence of any one researcher's subjective approach. Coding trees are typically developed through an iterative and adaptive process: as analysts work through a subset of interviews, they may revise, refine, or expand the coding scheme until a final structure is established. Intuitively, it is important that the number of documents reviewed in creating the coding tree is sufficiently big to capture the variation across the whole sample, so researchers generally increase this sample size until they feel that it is no longer necessary to add additional categories.

Once the coding tree is finalized, codes are applied at a standardized level—such as a line of text, a question–answer pair, or an entire interview. In our study, we assigned codes at the level of question-answer (QA) pairs to facilitate out-of-sample prediction. To further support this predictive analysis, the qualitative codes were assigned in a binary format. The coding tree and annotation process in our application to aspirations are described in Section 3.

2.2 Our supervised NLP approach

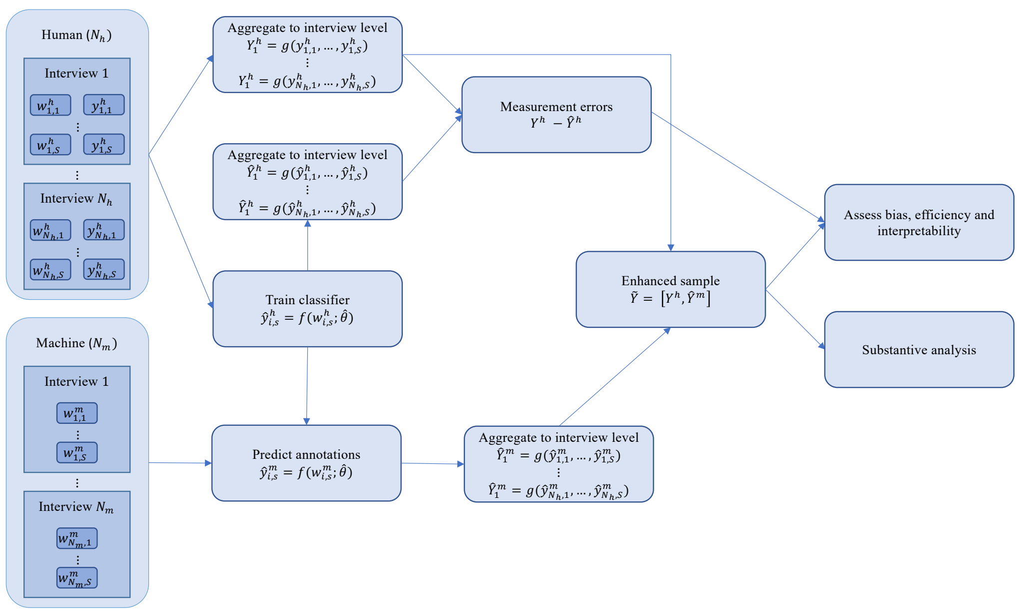
Our goal is to use our subset of annotated interviews to provide reliable annotations for the remainder of the sample, thus allowing statistical analysis that would not be possible on a smaller sample. We do this by training classifier models on our annotated set and then using this model to predict 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. As qualitative annotations are binary and defined at a standardised level (QA pairs) we can represent each code as an independent QA-level binary classification problem.[7]

Figure 1 illustrates our overall methodology for a single code. 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. We represent $w^h_{i,s}$ as the text of the sth QA in interview i in the human sample, with $y^h_{i,s}$ being the binary annotation of the code on that QA. In other words, $y^h_{i,s}$ will be equal to one if that QA has been annotated as displaying that code, and will be zero otherwise.

We then train a classifier $f()$ parametrised by θ to predict $y^h_{i,s}$ based on the QA text $w^h_{i,s}$. There are many options for both the classifier we can use here, as well as how to represent the text numerically (discussed in Section 2.3). A key point here is that the text representation is unsupervised - i.e. we do not use any information about y or about the interview subject

[7] In future work, we aim to explore the potential of using more nuanced coding schemes, including multi-label or probabilistic ("maybe") codes.

Figure 1: Summary of iQual methodology



Note: This Figure shows a visual summary of our methodology for creating an enhanced sample from a smaller set of human annotated interviews.

when creating a numerical representation of the text. Given this trained classifier we can then predict annotations $\hat{y}^m_{i,s}$ at the QA level for our unannotated "machine" sample.

Training at this granular level, rather than at the level of the whole interview has two advantages. Firstly it allows our qualitative coders to be more precise: 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.

The observed and predicted QA level annotations are aggregated to interview level with function $g()$. The choice $g()$ is in part a choice that depends on the research question. For example, the mean value of y across an interview's QA pairs gives a measure of the intensity with which this concept comes up, while the maximum value identified interviews where this concept comes up at least once. The observed human annotations and machine annotations are combined into the enhanced sample $\tilde{Y}$. We can use the enhanced sample for substantive analysis, taking advantage of the sample size to identify effects that would otherwise not be observable. The predicted annotations for the human sample can also be used to assess the measurement errors introduced by the model (see Section 2.4).

2.3 Modelling choices

Using the QA pairs to predict their interview-level annotations requires a numerical representation of the text, a classifier model and a aggregation function $g()$. While the choice of aggregation function is something that we leave to researchers' discretion, we use k-fold cross-validation to select the text representation and the classifier, as well as to tune any necessary hyperparameters. Training classifiers for each annotation independently allows for the fact that different classifiers or text representation may be optimal for different codes. The options we select over in our main results are discussed in Section 4.1, while Appendix C.1 gives the full list currently implemented in our Python package, and we describe how this package works in practice in Appendix E. We 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).

Although the main results we present in this paper are based on a training set of 400 interviews with a test set of a further 389, our cross-validation approach also allows us to approximate out-of-sample performance even when using the entire annotated sample as a training set. Our enhanced sample results in our application to aspirations are very similar whether we use 400 interviews or the full 789 as a training set, as shown in Appendix D.6.

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 7, we also train models on a subset of the training data which is sampled without replacement.

2.4 Assessing the enhanced sample's value

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 is an improvement. By quantifying the measurement errors in our test/validation sets and comparing results in the human and enhanced sample we can assess whether our enhanced sample adds value along two dimensions: bias and efficiency. The potential advantage of our method is that the larger sample size allows for statistical analyses that would not be possible in a smaller human annotated sample. For this advantage to be realised, it must be the case that the extra annotations do not introduce a bias into results and that the measurement error is small enough that it results in an increase in efficiency (i.e. more precise estimates).

Most straightforwardly, we can (and do) check the out-of-sample performance of our classifiers, in a held-out test set or validation sets. However, while this is certainly of interest and we need to verify that our classifiers perform better than random guesses, it is only a means to the end of a larger sample size for further analysis. We therefore care much more about whether the inevitable measurement error introduces a bias into the sample, or are so noisy that they lead to a decrease in efficiency (i.e. do not lead to reduced standard errors). We describe a test for each of these below.

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 a code for men and under-predict for women, 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. This is particularly important if we have reason to think that the human annotated sample is different from the unannotated sample in some important way.

We can test for potential bias by testing the association of test or validation set prediction errors with any explanatory variables of interest, which we will refer to for the rest of this Section as “interview characteristics.” These interview characteristics could be attributes of the interviewee (e.g. gender, age, education or occupation) or of the interview itself (e.g. when and where the interview occurred). A simple way to do this, as demonstrated in Section , is to regress the prediction errors on a range of explanatory variables of interest, with the F-statistic of these regressions forming a natural test of bias. If a bias is identified, this can be quantified and accounted for in subsequent analysis.

Efficiency. Even if measurement error does not introduce a bias in the machine annotations, it will add extra noise to these observations. How one assesses the efficiency gains of an enhanced sample will of course depend on the specific context and research question being considered. However, we propose a general test that can be used to demonstrate that an enhanced sample yields more precise results than the human annotated sample, while making minimal assumptions on the nature of the relationships in the data. This test compares the statistical significance of regressions of annotations on interview characteristics in the enhanced and human annotated samples.

Assuming text-based variables are related in some way to interview characteristics, then our enhanced sample ought to give stronger evidence of this relationship. We can therefore compare F statistics for regression of annotations on interview 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. Note that this test does not require any assumptions on how the text and interview characteristics are related, just the relatively weak assumption that there is some relationship between them.

An alternative, more stylised test of efficiency is discussed in Appendix C.3. Following Elbers *et al.* (2003), we show how we can account for two of the types of error in our machine annotations: prediction error and model error.[8] Assuming that these errors are independent and normally distributed, we can then directly quantify their effect on a statistic of interest. We show that the standard error on a sample mean estimated in the enhanced sample will therefore be smaller than that estimated in the human sample if the following condition is met:

$$\frac{\hat{\sigma}_m^2}{\hat{\sigma}_h^2} < \frac{N_m + 2N_h}{N_h}. \tag{1}$$

where $\hat{\sigma}_m^2$ is the variance of the machine annotated sample, accounting for prediction and model errors and $\hat{\sigma}_h^2$ is the variance of the human annotated sample. In words, the estimates will be more precise in the enhanced sample if the additional variance due to measurement error is low relative to the increase in sample size. We calculate these variances along with adjusted standard errors for a sample mean across the codes in our application. In all cases, the standard error of the enhanced sample mean is lower than that of the human sample (Table C.12), showing increased precision.

3 Empirical Example - Aspirations of Rohingya Refugees

In this Section we introduce the context in which we will apply our method - interviews with Rohingya refugees and their Bangladeshi hosts about their aspirations and ambitions for their children. First, we draw on theories from economics, anthropology and philosophy to distinguish between the concepts of ambition, aspirations and navigational capacity. We

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

then describe our data collection and qualitative analysis processes, as well as details of the quantitative variables available to us.

3.1 A conceptual framework

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 neighbourhood" 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 large empirical literature that analyses how aspirations matter in a variety of important spheres, and particularly in educational and labour market investments (Fruttero *et al.*, 2025).

The empirical literature is based on quantitative measures of aspirations using structured questionnaires and, perhaps consequently, does not delve into the broader dimensions of aspirations that Ray first highlight 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." This is a cultural and cognitive resource that allows them to navigate their way to a better future, that we refer to throughout as "navigational capacity."

Furthermore 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." They are the learning process involved in "becoming someone" - of acquiring norms, perspective, values around which "our lives will come to turn." An "ambition", on the other hand, is a specific goal that "she is fully capable of grasping in advance of achieving it" (Callard, 2018, page 229). Ambition is "directed at those goods – wealth, power, fame – that can be well appreciated even by those who do not have them." By Callard's definition, economists' understanding of aspiration is more in line with "ambition" rather than "aspiration", a distinction that we adopt in this paper.

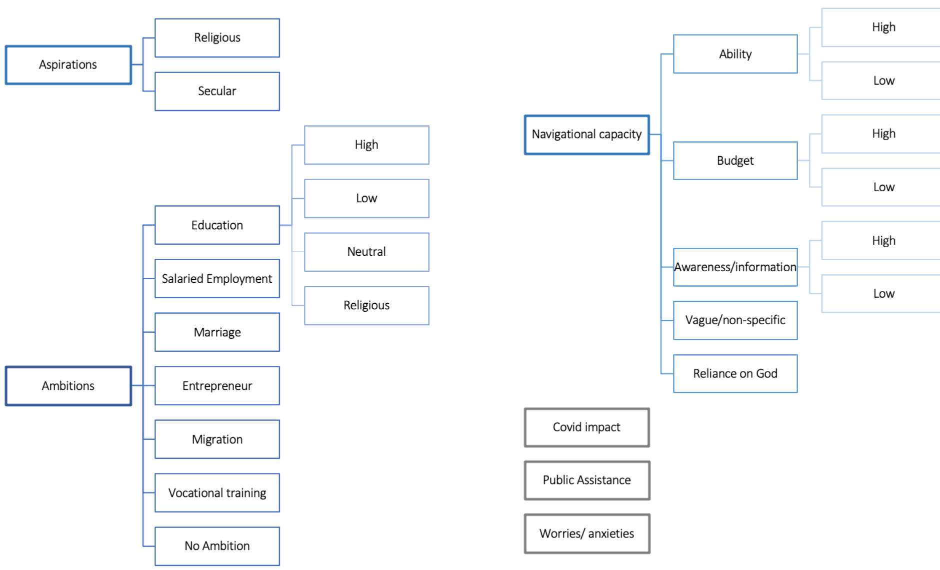
These distinctions are not just semantic. They have implications for measurement. Navigational capacity, being a cognitive and culturally determined capacity, is likely to be less amenable to structured questions where responses to questions are not easy to predict in advance. Similarly, aspirations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

These distinctions could have potentially important implications for policy – if navigational capacity matters for economic outcomes this could justify interventions to improve cognitive ability, as might interventions to guide less advantaged people towards achieving their goals. If aspirations have distinct determinants and effects from ambitions, 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. In measuring and quantifying these concepts, our approach allows us to take an important step towards potentially integrating them into the policy-making process.

3.2 Open-ended interviews

The data analysed in this paper are from Cox's Bazaar in Bangladesh where about 750,000 Rohingya refugees who were forcibly displaced from Myanmar between 2017-18 are primarily housed. We survey from both this population of Rohingya refugees and the local Bangladeshi host population. Our data comes from a three round survey. First, a baseline survey was conducted of 5,020 households in 2019 covering a wide range of topics (finances, education, gender, age, trauma) through traditional quantitative survey questions. There were then two further rounds in which we conducted open-ended interviews. "Round 2" in 2020 included ∼10 minute interviews on aspirations with household head of 1,020 of the previously sampled households. "Round 3" was conducted in 2021 and included a further 2,040 open-ended interviews with household heads, this time ∼30 minutes long and covering three distinct topics: aspirations, belonging and well-being. In this paper, we focus on the interviews about aspirations. Of the 3,038 interviews conducted we restrict ourselves in this analysis to households that whose eldest child lived with them and was still of school-going age. which leaves us with 2,177. Further detail on the Cox's Bazaar context and our data collection are in Appendix A.

Figure 2: Coding tree



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 machine translations of the transcripts that were manually corrected.[9] Across both Round 2 and Round 3, the interviews are on average 12.6 QA pairs long, with the average being 13.7 words long. In addition to the open-ended interviews, we also use several quantitative variables from the baseline survey on household characteristics. Table 2 shows summary statistics for these variables. In Round 2 subjects were also asked to rate their education ambitions on a scale from 1 to 7, where 1 is no education and 7 is postgraduate.

[9]We discuss the merits of machine vs human translation in more detail in Section 2.

3.3 Qualitative coding

We follow a traditional “flexible” coding processes (Detering and Waters, 2018), which combines careful inductive reading of the qualitative interviews within a conceptual framework that combines theoretical insights from economics (Genicot and Ray, 2020), anthropology (Appadurai, 2004) and philosophy (Callard, 2018). We develop a “coding tree” which classifies interview transcripts at the QA pair level across 25 hierarchical and potentially overlapping categories.[10]

Figure 2 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 seven major categories – education (further sub-coded into High, Low, Neutral and Religious), Salaried Employment, Marriage, Entrepreneurship, Migration, Vocational Training, and No Ambition. Navigational Capacity was coded into six sub-codes – Low and High “Ability” and Low and High “Budget”. There were three additional codes that did not fit into the structure of aspiration, ambition and navigational capacity (Covid Impacts, Public Assistance and Worries/Anxieties). Descriptions and examples of each code are displayed in Appendix B.3.

[10]Further details on the development of the coding tree and coding process are given in Appendix B.

Table 1: Qualitative human annotations summary

Category	Code	Proportion of QA pairs R2	R3	Total	Proportion of interviews R2	R3	Total
Aspirations	Aspiration: Religious	0.036	0.028	0.030	0.208	0.234	0.221
	Secular	0.053	0.028	0.036	0.332	0.332	0.332
Ambition	No Ambition	0.009	0.012	0.011	0.056	0.137	0.096
	Salaried Employment	0.089	0.107	0.102	0.456	0.731	0.593
	Vocational Training	0.007	0.004	0.005	0.041	0.041	0.041
	Entrepreneur	0.031	0.012	0.018	0.154	0.122	0.138
	Education High	0.064	0.048	0.053	0.375	0.454	0.414
	Education Neutral	0.185	0.062	0.101	0.772	0.574	0.673
	Education Low	0.013	0.035	0.028	0.094	0.475	0.284
	Education Religious	0.035	0.016	0.022	0.210	0.168	0.189
	Marriage	0.082	0.036	0.051	0.385	0.396	0.390
	Migration	0.022	0.007	0.012	0.104	0.079	0.091
Capacity	Vague Non Specific	0.066	0.017	0.033	0.420	0.234	0.327
	Reliance On God	0.039	0.017	0.024	0.243	0.228	0.236
	Ability High	0.048	0.032	0.037	0.311	0.424	0.368
	Ability Low	0.035	0.033	0.034	0.230	0.360	0.295
	Budget High	0.033	0.013	0.020	0.200	0.188	0.194
	Budget Low	0.111	0.039	0.062	0.522	0.401	0.461
	Awareness Information High	0.070	0.031	0.043	0.387	0.297	0.342
	Awareness Information Low	0.007	0.012	0.010	0.061	0.145	0.103

Note: This Table shows the prevalence of each code across human annotated interviews in each round and across the whole human annotated sample. The first three columns show the proportion of QA pairs in which the code occurs, while the last three columns show the proportion of interviews in which the code appears at least once.

Using Atlas-TI, a qualitative data analysis software, a team of two co-authors (AH and AK) coded 789 transcripts. 400 interviews (comprising 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 interviews, 11 were left uncoded due to either poor audio, call drops-off, or very short responses with no plausible code applicability. Coders annotated interviews at the QA pair level to preserve granularity while being able to replicate the sub-division of interviews in the unannotated documents. We follow a standard approach to ensuring cross-coder agreement, with each interview being reviewed and any disagreements resolved through discussion between AH, AK and MB (Appendix B.2).

Many annotations are very sparse at the QA pair level, as illustrated by Table 1 which shows some summary statistics for the human annotations. For example only 3.0% QA pairs are annotated as Religious Aspirations. However, when aggregated to the interview level there is much less sparsity (for example, 22% of interviews have at least one Religious Aspirations annotation). Although the same coders and coding tree were used throughout, there are also notable differences between rounds. These are partly due to slightly longer interviews in Round 3 as well as changes in circumstances/attitudes over the intervening year.

3.4 Quantitative variables

In addition to the open-ended interviews, we also use several quantitative variables from the baseline survey. Table 2 shows summary statistics for these variables. In addition to the open ended interviews, subjects in Round 2 were asked to rate their education ambitions on a scale from 1 to 7, where 1 is no education and 7 is postgraduate level. This variable can also be used as a quantitative measure of education ambition, as discussed in Section 6.1. Going forward, when we refer to “household characteristics” we mean the variables in Table 2, with the exception of Quant Education Ambition.

4 Methodological Results

In this Section, we provide details on the model training process and which supervised models, text representations were selected. We discuss out of sample performance of our

Table 2: Quantitative variable summary statistics

Variable	N	Mean	St. Dev.	Notes
Refugee status	2,407	0.460	0.498	Dummy variable, 1 for refugee
Female eldest child	2,293	0.474	0.499	Dummy variable, 1 for female
Eldest child's age	2,295	10.381	4.636	Integer for age in years
Female household head	2,407	0.185	0.388	Dummy variable, 1 for female
Household head's age	2,407	38.726	11.768	Integer for age in years
Number of children	2,405	2.675	1.463	Integer for number of children
Parent's years of education	2,398	3.548	3.837	Integer for number of years
Parent's religious education	2,398	0.036	0.186	Dummy variable, 1 for religiously educated
HH asset index	2,406	0.147	1.820	Principle Component of assets owned following Filmer and Pritchett (2001)
HH income	2,407	1.125	2.340	Last month's income (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

Note: This Table shows summary statistics for the quantitative household characteristic variables used throughout.

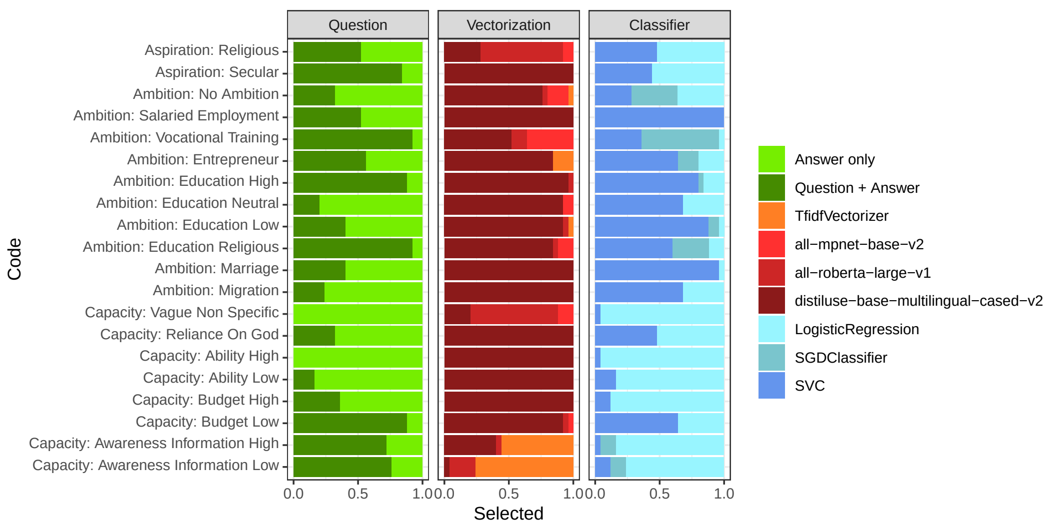
classifiers, before testing whether our enhanced sample has introduced bias and/or increased efficiency. Our results in this Section and the next are based on a training set of 400 interviews and so a held-out test set of 389 interviews. In Appendix D.6 we show results where the full 789 interviews were used as a training set and validation set errors are used in place of test set errors.

4.1 Model Training

We have 9,964 distinct QA pair level observations in the human annotated sample, which come from the 789 annotated interviews. In our main results, we use a subsample of 400 interviews as our training set and the remaining 389 as a held-out test set. This training set is randomly sampled, but weighted so that it is balanced with respect to refugee status and eldest child's gender. We select the text representation, classifiers and hyperparameters through three-fold cross validation. We re-train the models 25 times with a different draw for the validation set split and any stochastic processes involved in training the model, to provide a bootstrapped measure of the uncertainty in our predicted annotations. In Section 7 we show how varying the size of the annotated sample affects performance.

We allow the text representation to vary along two dimensions, illustrated by the first two panels of Figure 3, which show the proportion of bootstraps in which each representation is chosen. First, we include the answer of each QA pair only, or both the question and answer. As shown in the first panel of Figure 3, in many cases only the answer is selected. However, for some annotations, both the question and answer are usually included. Second, 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 C.1. When using count based metrics we allow for both single and two word phrases. In our main results, we use machine translations (Google Translate) of the Bengali transcripts into English. As discussed in Appendix C.2, we found that the machine English translation outperforms human translation in most cases as well as Bengali and Latin-transliterated Bengali. This may be because the translation process reduces noise while preserving the signal.

Figure 3: 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 the selected vectorizer, which is applied to convert the text into numeric vectors. The third panel shows the selected classifier.

We select over a range of five popular classification models including logistic regression, random forest, support vector classifier, stochastic gradient descent classifier and a simple neural network, see Appendix C.1. Each of these models have a separate set of hyperparameters that are chosen through 3-fold cross validation. 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. We do this by computing the mean annotation value across QA pairs within an interview, which we refer to as the "prevalence" of this code. In the notation of Section 2, this prevalence measure for a given code is defined as

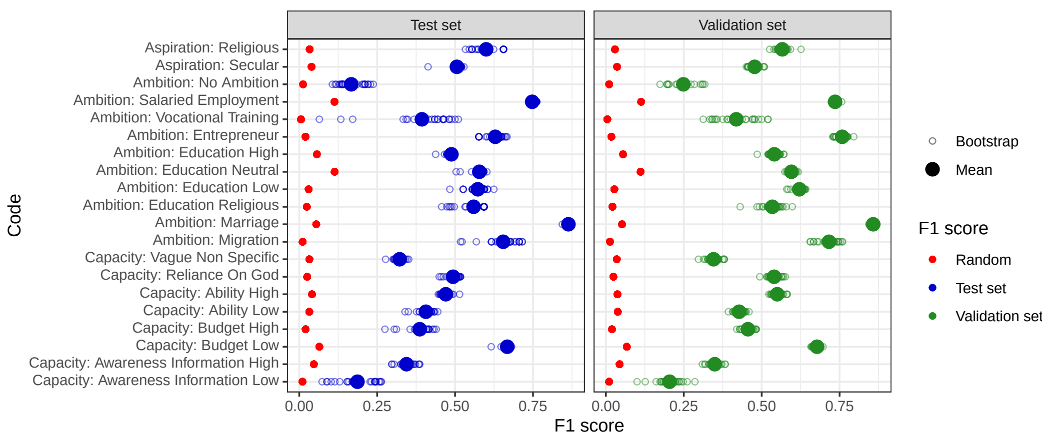
$$Y_i = g(y_{i,1}, ..., y_{i,S}) = \frac{1}{S}\sum_{s=1}^{S} y_{i,s} \tag{2}$$

where $y_{i,s}$ is the binary annotation for QA pair s in interview i. 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.

4.2 Out-of-sample performance

In our application, we have a large enough set of human annotated transcripts that we can separate these into a test and training set, however in other applications this may not be possible. In such a situation validation set performance can still gives us a measure of how the classifiers perform out-of-sample. We therefore present results here both in our test set and the validation set to illustrate that in a more resource constrained environment, researchers would still be able to get a sense of their models' out-of-sample performance even without a dedicated test set. The analysis in Section 7 also confirms that validation set performance is a good indication of out-of-sample performance. Given that many of our codes are very sparse, a raw accuracy statistic (i.e. proportion of correctly classified observations) is not a good measure of performance. Instead, we use F1 scores which balances precision (ratio of true positive predictions to total predicted positives) and recall (ratio of true positive predictions to total actual positives).[11]

Figure 4: Out-of-sample prediction performance



Note: This Figure shows test set (left panel) and validation set (right panel) performance of the selected model for each code and each bootstrap run, as measured by the F1 score. The sparsity of the codes 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 4 shows the test and validation set F1 scores, across each code and bootstrap run. Sparsity, which corresponds to the expected F1 score under random guesses, is shown in red as a natural benchmark. In all cases our text-based models do much better than random

[11]As shown in Figure D.2, because of the sparsity of the annotations, accuracy in all cases is over 90%. Precision and recall scores for each code are roughly balanced indicating that there is neither over- or under-prediction.

and validation set performance is an excellent predictor of test set performance. However, there is considerable heterogeneity across codes. In particular, codes associated with less concrete concepts (e.g. Awareness Information Low, No Ambition and Vague Non Specific) appear to be harder to predict.

A key question we explore in more detail in Section 7, is how many human annotated interviews is "enough". Increasing N_h, the number of human annotated interviews, will improve 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 annotation quality is both time consuming and expensive. To assess how much performance improves as the annotated sample size N_h increases, we draw subsets of our human annotated sample without replacement, varying the sample size from 100 to the full 789 (drawing 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.

As N_h increases, there are improvements in both validation set and test set performance (see Figure D.6). When the human annotated sample is very small, test set performance is worse than validation set performance, but from a training set of around 400 interviews performance converges for most codes. 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 appear to be diminishing returns.[12]

[12]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.

4.3 Bias and efficiency tests

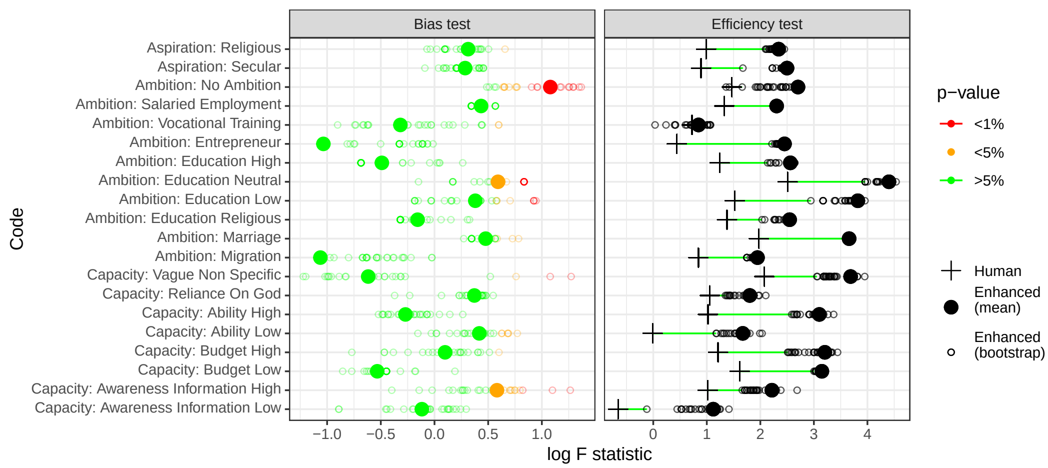
In Section 2.4 we proposed two tests to assess the value of our enhanced sample, in the context of the inevitable measurement errors, along two dimensions: bias and efficiency.

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 certain codes 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. We test the association of prediction errors with household characteristics for each interview described in Table 2. We use the test 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.

The left panel of Figure 5 shows the F-statistics for this bias test across each code. The test is carried about for each bootstrap iteration (show as the hollow points) as well as for the median 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 one case there is evidence of at the 1% level of a relationship between household characteristics and prediction errors - No Ambition. In a further two cases there is evidence at the 5% level - Awareness Information High and Education Neutral, so we can inspect these in more detail. The regressions in question, shown in Table D.13, indicate that No Ambition is slightly over-predicted for households with an older and male household head, as well as slightly over-predicted for parents of a female eldest child. For Education Neutral,

Figure 5: Bias and Efficiency tests



Note: This Figure shows the results of our bias and efficiency tests. For both panels, the hollow circles represent the results for each bootstrap run and the solid circles represents results based on the mean prediction across each bootstrap. The left panel shows the results of our bias test - the log F statistic for the regression of the test set errors on household characteristics, for each code. The colour of each point indicates the significance level of the F statistic. The right panel shows the results of our efficiency test - the log F statistic for the regression of each code on household characteristics in the enhanced and human samples. The human sample statistics are shown as crosses and the enhanced sample as circles. In every case the statistic increases in the enhanced sample, suggesting an increase in efficiency.

we see over-prediction in interviews that where carried out in Round 3 (i.e. in 2021 rather than 2020).[13] These effects are quantitatively quite small, but we will need to bear them in mind in our substantive analysis. We also include a dummy variable for whether an interview is machine or human annotated in any regressions using enhanced sample data, and show results for the human sample in parallel to enhanced sample results.

Efficiency. As described in Section 2.4, we use a general test of whether our enhanced sample creates efficiency gains. This test compares the F statistics of regressions of annotations on household characteristics in the enhanced and human annotated samples. If the enhanced sample increases this F statistic relative to the human sample it suggests that the larger sample leads to more precise and interpretable results in spite of the additional measurement error.

The right panel of Figure 5 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, although in some cases this difference is quite small (e.g. Vocational Training). So while there is no theoretical guarantee of increased precision and interpretability when we enhance our sample, in all our cases we see an increase for the mean across bootstrap iterations. This is confirmed in Section 5, where we find increased precision in the enhanced sample in virtually all cases.

[13]Results for some alternative versions of the bias test are shown in Figure D.3, which is based on the prediction errors in the training set and Figure D.14 which is based on a specification in which the full annotated sample is used as a training set. Although the precise details of the biases vary across specifications, the overall picture is quite similar.

5 Substantive Results

In this Section we highlight our key results on ambition, aspirations and capacity, based on our data from Cox's Bazaar. Subsection 5.1 analyses the determinants of open-ended and quantitative responses on ambition. Our findings on ambition are largely consistent with previous literature: refugees, parents of female children and less educated parents generally display lower levels of ambition. We also identify interesting correlations between Callardian aspirations and ambition. In Subsection 5.2, we analyse our open-ended measures of secular and religious aspirations. We find that aspirations show distinct gender effects with parents of female children less likely to express both religious and secular aspirations. We also see that more educated parents' aspirations are more likely to be secular. Finally, Subsection 5.3 explores the variation in our open-ended measures of navigational capacity for Rohingya refugees and Bangladeshi hosts. Here we find that, refugees are less likely to display low navigational capacity across several dimensions, which we argue could be evidence of a novel migrant selection effect.

Across this Section we show results for the enhanced and human samples in parallel to illustrate that the statistics estimated on the enhanced sample generally have similar point estimates to their human sample counterparts, but are considerably more precise. This tighter standard errors allow us to draw conclusions in the enhanced sample that would not be possible from the human sample alone.

5.1 Ambition

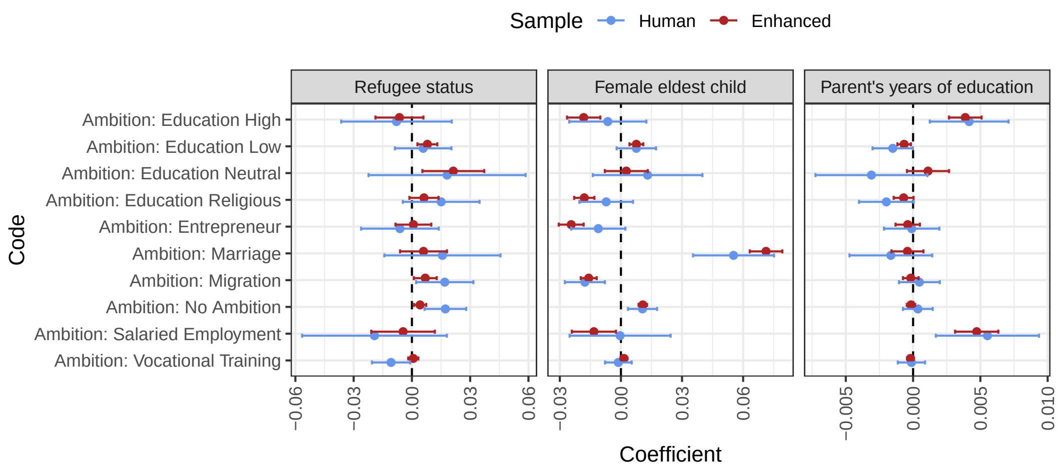
Our coding tree (Figure 2) includes seven categories for ambition. The Education category is split into four sub-codes (low, neutral, high and religious). The remaining six categories refer to specific material goals: Entrepreneur (non-wage enterprise job); Marriage; Migration (leaving the current place of residence for work, study or settlement); Salaried Employment (specific salaried occupations); Vocational training (vocational rather than academic training) and No Ambition (expressions of helplessness, implied unwillingness to or lack of dreaming and planning).[14] As this coding structure was developed after reading the transcripts, the categories arose from the topics that our interview subjects talked about, rather than being imposed by researchers beforehand. These dimensions of ambition are thus the dimensions that are important to our interview subjects.

Table 1 shows that most parents express clear "ambitions" with 60% wanting their children to have salaried jobs, while 14% would like them to be entrepreneurs. Education, consistent with the literature on aspirations, is also a priority with 41% wanting a "high" education for their children (defined as post high school), 19% want them to have an Islamic religious education, and 39% would like to see them married.

The prevalence of these different dimensions of ambition varies considerably across households. Figure 6 displays coefficients and confidence intervals from regressions of the prevalence of these codes on household characteristics for human and enhanced samples. The key methodological point is that the point estimates of coefficients from the human and enhanced samples are generally close together, but the enhanced sample coefficients have smaller standard errors.

[14]See Appendix B.3 for more detailed descriptions of each code.

Figure 6: Ambition and household characteristics



Note: This Figure shows coefficients for regressions with the prevalence of ambition codes on the left hand side and a range of household characteristics and other control variables on the right hand side. Coefficients estimated on the human coded sample are shown in blue, with the estimate using the enhanced sample in red. We display coefficients on refugee status, female eldest child and parent's years of education here, in separate panels from left to right. The underlying regressions contain additional household characteristics: number of children, whether household head is female, age of household head, whether parent is religiously educated, age of eldest child, household assets, household income and a parent trauma score. We also include an indicator for which round the interview was conducted in and for the enhanced sample an indicator for whether that transcript was machine annotated. Full results for these regressions are shown in Appendix D.5.2.

Looking at Figure 6, we see a clear relationship between gender and ambition. Consistent with the existing aspirations literature (Beaman *et al.*, 2012; Favara, 2017; Roy *et al.*, 2018), parents have higher ambitions for male over female children. Parents of female children express lower ambition on salaried employment, migration, entrepreneurship, education and religious education, and are more likely to express No Ambition for female children. They are also, overwhelmingly, more likely focus their ambitions for female children on marriage. Also consistent with the existing literature, more educated parents are more likely to have high education ambitions for their children and also have ambitions that are directed towards salaried employment. Rohingya and Bangladeshi parents show similar levels of ambition on most occupation and education levels.

Table 3: Quantitative ambition variable regressions on enhanced sample

Dependent Variable:	Quantitative education ambition (scale from 1 to 7)			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	4.0*** (0.18)	4.1*** (0.20)	3.7*** (0.36)	3.7*** (0.40)
R3	-0.001 (0.08)	-0.02 (0.18)	0.06 (0.16)	0.04 (0.18)
Refugee status	-1.5*** (0.12)		-1.5*** (0.20)	-1.5*** (0.21)
Number of children	-0.01 (0.03)		0.06 (0.06)	0.05 (0.06)
Female HH head	0.02 (0.10)		0.05 (0.19)	0.11 (0.19)
Age of HH head	0.008* (0.004)		0.007 (0.007)	0.007 (0.007)
Parent's years of education	0.08*** (0.01)		0.04* (0.02)	0.03 (0.02)
Religioulsy educated parent	0.74*** (0.20)		0.42 (0.34)	0.41 (0.34)
Female eldest child	-0.38*** (0.08)		-0.15 (0.14)	-0.12 (0.15)
Age of eldest child	0.03** (0.01)		0.03 (0.02)	0.04 (0.02)
HH asset index	0.06** (0.03)		0.05 (0.05)	0.05 (0.06)
HH income	0.02 (0.02)		0.009 (0.03)	0.01 (0.03)
Parent trauma experience	0.03 (0.02)		0.01 (0.03)	0.02 (0.03)
Machine annotated		-0.07 (0.15)	-0.11 (0.13)	-0.09 (0.13)
Ambition: No Ambition		-10.9*** (2.8)	-4.3* (2.4)	-4.1* (2.4)
Ambition: Salaried Employment		2.7*** (0.64)	1.7*** (0.57)	1.7*** (0.60)
Ambition: Vocational Training		-3.8* (2.0)	-2.4 (1.7)	-2.8 (1.7)
Ambition: Entrepreneur		0.52 (1.0)	-1.7* (0.95)	-1.5 (0.96)
Ambition: Education Low		-4.1*** (1.6)	-1.2 (1.4)	-1.8 (1.4)
Ambition: Education Neutral		-0.88* (0.50)	0.03 (0.44)	-0.27 (0.46)
Ambition: Education High		3.6*** (0.82)	2.6*** (0.75)	2.4*** (0.77)
Ambition: Education Religious		-3.3*** (0.97)	-1.4* (0.85)	-0.36 (1.2)
Ambition: Marriage		-1.9** (0.72)	-1.8*** (0.69)	-1.9*** (0.69)
Ambition: Migration		-0.06 (1.3)	1.1 (1.4)	1.7 (1.4)
Aspiration: Religious				-1.3 (1.1)
Aspiration: Secular				1.5** (0.67)
Capacity: Vague Non Specific				0.40 (0.83)
Capacity: Reliance On God				-0.55 (0.94)
Capacity: Ability High				2.4** (1.1)
Capacity: Ability Low				0.65 (1.1)
Capacity: Budget High				-1.5 (1.3)
Capacity: Budget Low				-0.85 (0.65)
Capacity: Awareness Information High				-0.58 (0.72)
Capacity: Awareness Information Low				2.0 (2.1)
Fit statistics				
Observations	1,183	426	392	392
R^2	0.396	0.286	0.519	0.538
Adjusted R^2	0.390	0.266	0.489	0.495
F-test	64.1	13.8	17.3	12.6

*IID standard-errors in parentheses, Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: This Table shows results for regressions of the quantitative education ambition variable on household characteristics and the qualitative measures. Results here are only for the enhanced sample, with human sample results shown in Table D.22.

Parents ambitions for their child's education is has been widely studied in the aspirations literature using standard quantitative survey questions. Our data includes such a question, as described in Section 3, which allows us to compare our qualitative variables with this quantitative one. Table 3 which shows results regressing the same set of household characteristics on the quantitative education response, and then adds the qualitative ambition codes.

The quantitative and qualitative education ambition results have very similar signs and significance levels, but the qualitative measures add important context. The refugee coefficient on quantitative 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 (as shown in Figure 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 no explicitly less ambitious on education than hosts, but that having had disrupted lives they have more trouble expressing a clear education ambition. Column (3) of Table 3 also shows that when the qualitative education ambition codes are added to the regression on quantitative education ambition, they add explanatory power without substantially changing the coefficients on the original set of right hand side variables suggesting that are capturing different dimensions.

While ambition is relatively easy to elicit using structured questions, latent concepts like "aspiration" and "capacity" are harder because they are more subtle. We show that an open-ended discussion on aspirations and navigational capacity captured a great deal of content. We see in column (4) in Table 3 that these concepts have meaningful correlations with quantitative education ambition. Parents who value secular aspirations for their children tend to have higher education ambitions, while high ability parents also tend to express higher levels of education ambition.

5.2 Aspirations, gender and education

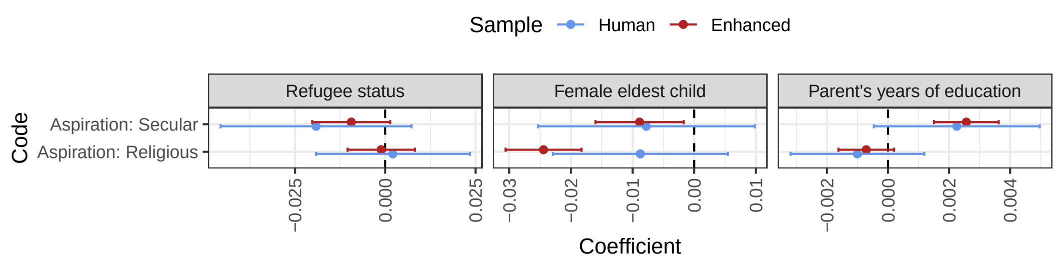
Aspirations are a cultural, normative process which we have coded as religious and secular aspirations for children. We observe both in a high proportion of interviews, 22% of respondent demonstrate religious aspirations for their children and 33% express secular aspirations (Table 1).

A QA pair is tagged as "Aspirations: Secular" when a parent expressed goals 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 will attain decent standard of living. Examples here include: taking care of wife and children and elderly parents; earning enough to live a beautiful life; being healthy and having a respectable job; being given recognition; earning well and building a house; being happy; and becoming a doctor for the good of the nation.

"Aspiration: Religious", on the other hand, is coded when a parent expressed an aspiration (rather than an ambition) for their child that was religiously motivated. Examples that are coded for religious aspirations include: ability to read The Quran;

maintaining Islamic dress; praying regularly; working in an Islamic bank; becoming a maulvi/alem/alemdar/elamdar/mawlana (i.e an Islamic scholar); becoming hafiz/hafez or attending hafez khana (i.e memorize Quran); attending noorani madrassa or madrassa school (i.e schooling for religious education); and learning/studying Arabic. Pinning down these nebulous concepts is difficult, and so challenging in a quantitative survey.

Figure 7: Aspiration and household characteristics



Note: This Figure shows coefficients for regressions with the prevalence of aspiration codes on household characteristics. A full description of the figure is given in the note for Figure 6 and the full regression results are shown in Appendix D.5.3

Figure 7 displays coefficients and confidence intervals from regressions of the prevalence of these codes for human and enhanced samples. We see that "aspirations" also show distinct gender effects. Parents of female children are less likely to express both religious and secular aspirations for them reserving those more moral desires for their male children. Rohingya refugees and their Bangladeshi hosts have similar conditional levels of religious and secular aspirations. However, educated parents are far more likely to express secular aspirations for their children rather than religious aspirations.[15] This suggests that the lack of inter-generational mobility observed in education (Black *et al.*, 2005; Black and Devereux, 2011) may also extend into the inter-generational transmission of norms and values. Callardian

[15]Note that several of these effects are only statistically significant in the enhanced sample.

aspirations can be distinguished and measured apart from goal-oriented ambitions should perhaps be explored further, moving beyond the focus on goal-oriented ambitions that the empirical literature on both aspirations and inter-generational mobility currently has. Open-ended questions have the potential to explore such concepts that reflect different disciplinary conceptions about what aspirations mean.

Table 4: Definitions and Examples from transcripts of Navigational Capacity

Code	**Definition and examples from transcripts**
Vague/Non Specific	When parent mentioned unspecific or unclear attempts/measures to help achieve dreams for child. Examples include: trying hard; will do as much as I can; will do my best; let's see what happen.
Reliance on God	When either the parent fully/partially relies on God to fulfil future dream for children or is fully/partially reliant on God at present. Examples include: even if there is hope, it depends on God willing; god is running our lives somehow.
Ability High	Coded when the parent demonstrates having gone the extra mile ensure a better future for the child. Examples include: 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.
Ability Low	Coded when the parent specified having no resources to help the child. Examples include: what can we do from here? We are have to stay how we are.
Budget High	Coded when the parent expresses having money, including an ability to save/spend.
Budget Low	Coded when the parent expresses not having money.
Awareness Information High	Coded when the parent displays awareness of information relevant to their goals.
Awareness Information Low	Not knowing what to do, cluelessness.

Note: This Table shows the definition, and in some cases examples from the transcripts, to illustrate the meaning of Navigational Capacity codes. More details are given in Appendix B.3.

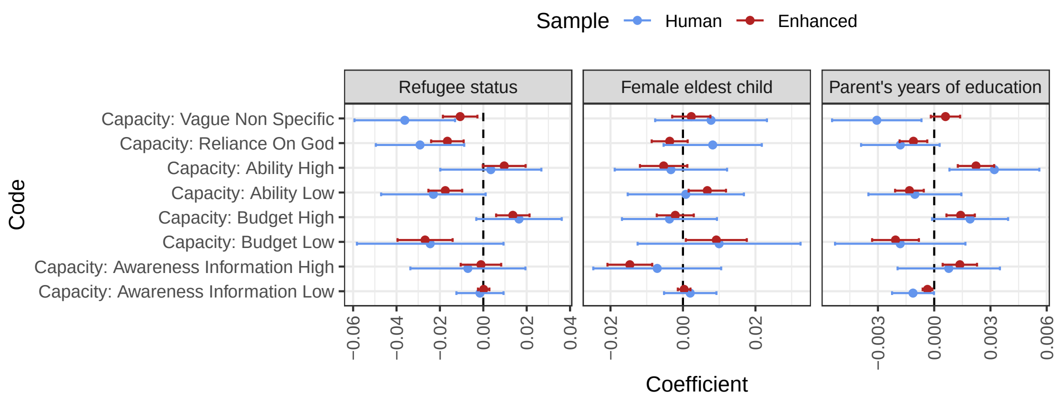
5.3 Navigational capacity, refugee status and education

Our open-ended interviews and interpretative qualitative structure can also quantify Appadurai's (Appadurai, 2004) notion of "navigational capacity"; a culturally shaped ability to imagine a better future and take concrete actions towards realizing it. Our coding tree, developed inductively through a close reading of the transcripts, quantifies navigational capacity across several different dimensions. Definitions and examples of each of these are shown in Table 4. The codes Ability High, Ability Low, Vague/Non Specific and Reliance on God are, in particular, directly related to the "navigational capacity". The codes Budget High and Low and Awareness-Information High and Low are more instrumental and straightforward.

Figure 8 displays coefficients and confidence intervals for these ambition variables, and focuses on the same household variables as previous Subsections. On child's gender, we find little significant variation in capacity, which is perhaps unsurprising as capacity is an attribute of parents rather than children. However, parent's education shows some striking patterns - more educated parents are less likely to be budget constrained. Importantly, parent's with low levels of education show low navigational capacity through a higher reliance on god, lower ability and lower information awareness. On the other hand, parents with high education are more likely to show high ability and high information awareness. This has clear policy implications and highlights a potential mechanism preventing inter-generational mobility. Low education levels associated with poverty may result in worse outcomes of children because parents have lower capacity to realize their aspirations for their children, in addition to lower budgets.

Interestingly, while refugees are considerably less educated than hosts (average of 1.94 vs 4.92 years of schooling) we consistently find that, conditional on our household variables,

Figure 8: Capacity and household characteristics



Note: This Figure shows coefficients for regressions with the prevalence of aspirations codes on household characteristics. A full description of the figure is given in the note for Figure 6 and the full regression results are shown in Appendix D.5.4.

refugees are less likely to display low capacity than hosts. Rohingya refugees give fewer vague responses about their plans to achieve their aspirations, are less likely to rely on god, and are less likely to display low ability. Although investigating the mechanism behind this is beyond the scope of this paper, this result could be explained by the fact that refugees who have managed to reach the camps in Bangladesh are a selected population who have survived a very difficult ordeal to get there. This would adds a new dimension to the literature on migrant selection (Chiswick (1999) which focuses on education and earnings with an active debate on whether migrants display positive or negative selection (e.g. (Chiquiar and Hanson, 2005), Borjas *et al.* (2019), Birgier *et al.* (2016)). Migrants, particularly refugees fleeing violence and conflict, may have greater navigational capacity than hosts given the challenges they have overcome in reaching the relative safety of a refugee camp.

Open-ended responses can thus give novel and meaningful measures of subtle concepts like Appadurai's navigational capacity and Callard's normative notions of aspirations, while being consistent with with earlier empirical work on concepts like goal-oriented ambition that have been extensively studied with structured questions.

6 Comparing to alternative methods

In this Section we compare our approach to alternative methods. We first discuss traditional quantitative survey methods in Subsection 6.1, we then consider whether Large Language Models (LLMs) might act as a substitute for our approach in Subsection 6.2 and then with then briefly turn to a range of other alternative NLP approaches to extracting information from text in Subsection 6.3. We find that some of these alternative approaches can have dangerous flaws (dictionary methods in Section 6.3.2 and the LLM annotation in Section 6.2), while other methods yield some interesting results, but do not give us clear measurements of aspirations or navigational capacity.

6.1 Quantitative survey questions

While open-ended interviews are relatively rare in economics, structured surveys and questionnaires are very common. In contrast to traditional quantitative survey questions, open-ended questions follow a more respondent-driven, "bottom-up" approach to eliciting responses, giving respondents the opportunity to engage in more natural conversation that allows less concrete and emergent concepts to be probed. This has the added advantage of allowing respondents more freedom to interpret complex concepts from their point of view rather than select a limited set of structured responses that reflect the researcher's interpretation. The emphasis on structured questionnaires is arguably one reason why the empirical aspirations literature in economics has focused on goal-oriented ambition rather than moral and normative aspirations, and culturally and cognitively driven ideas like "navigational capacity."

A good structured questionnaire requires a great deal of work to get right. For example Bernard and Taffesse (2014) describe how designing an aspirations questionnaire focused on goal-oriented ambition required repeated rounds of pre-testing and contextualization to develop. Indeed this process of developing a good structured questionnaire can involve open-ended interviews or focus groups in earlier rounds, as discussed by Jayachandran *et al.* (2021). Our experience in collecting these data has taught us that open-ended questions, by not requiring researchers to thoroughly anticipate a full set of possible of possible responses, can be designed with a shorter period of pre-testing. Furthermore, by allowing for a process of iterative questioning they allow interviewers to probe concepts that a structured interview may find difficult to capture altogether. However, they also require a very different approach to selecting and training interviewers who skills need to include the ability to engage in empathetic and informed conversations. This is particularly important for vulnerable or displaced populations where repeated sampling can be difficult and for whom surveys that work on other populations might not transfer well.

A question that arises with this is if responses to open-ended and structured questions are different? Are "facts" elicited with open-ended responses reliable? We can see an answer this in Table 3, which shows that our quantitative and qualitative education ambition variables display very similar patterns across our sample, but the qualitative question adds context and captures additional variation. Open-ended responses thus provide factual answers that are similar to those elicited from standard structured questions, but with more nuance.

Furthermore, more subtle concepts, like moral aspirations or navigational capacity, are hard to turn into good structured questions. This can be for a range of reasons: respondents have a heterogeneous understanding of the concept; responses can be difficult to predict making probes of multiple choice options hard to anticipate in advance; the concept in question may be something that the respondent themselves is not consciously aware of; or the complexity/abstractness of any quantitative question may make it difficult to ensure honest, thoughtful and conscientious reporting. In such cases, it may be possible to elicit responses that are a closer reflection of the respondent with an open-ended approach. In fact, in the case of aspirations and navigational capacity, we are not aware of any work that has attempted to quantify these in structured questionnaires which can be interpreted as evidence for the difficulty of formulating such structured questions.

6.2 Annotating with LLMs

An alternative way to scale up human coding to a larger sample would be to use a Large Language Model (LLM) for annotation, once a coding tree has been developed as described in Section 3. Previous work has shown that using these techniques can help LLM outperform crowd workers in text annotation tasks (Gilardi *et al.*, 2023). We can thus think of using LLMs for large scale annotation as conceptually similar to using untrained human annotators. We find that LLMs perform considerably worse than our "small" supervised models in terms of prediction performance, and also introduce a significant amount of bias through their prediction errors.

To test this, we created detailed textual instructions for each code, similar to those one would give to human annotators. We then provided these to several LLMs as prompts including precise directives for the LLM, employing "few-shot learning" (Brown *et al.*, 2020) by giving several examples, and "chain of thought" prompting (Wei *et al.*, 2022) by asking the LLM to explain the reasons for its decision. An example of this prompt is shown in Figure D.5. We tested three different LLMs: the closed-source ChatGPT (gpt-3.5-turbo) by OpenAI, as well as two open-source LLMs by Meta, the Llama-2 (13b) and its fine-tuned "chat" variant (Touvron *et al.*, 2023).

We evaluated the the prediction performance of the three LLMs compared to our method (iQual) in terms of accuracy, F1 score, precision and recall (see Figure D.4). All three LLMs perform worse in most cases across all metrics except for the recall score. The relatively good performance in terms of recall is because the LLMs tend to heavily over-predict most codes and so have fewer false negatives (but many more false positives). This over-prediction is a potentially dangerous source of bias that would need to be corrected for if researchers want to, for example, compare the prevalence of two different codes. Furthermore, if we run the bias test as described in Section 2.4 we find that there is much stronger evidence of bias with respect to household characteristics in the annotations provided by the LLMs than for those generated with iQual. This comparison between human, iQual and LLM annotations with a particular focus on bias is explored in more detail in Ashwin *et al.* (2025), in which we show that using LLM annotation in our dataset can lead to misleading inferences due to the significant bias that the LLM annotations introduce.

The crucial conceptual difference between LLMs and our methodology is that iQual is a "small" language model trained only on our expert human annotations while LLMs are trained on huge datasets of text from a wide range of different contexts. This may explain why iQual performs better in the case of our transcripts. Our data is drawn from a particular context and we want to extract a very specific set of features. By creating a bespoke training set through expert human annotation and then training much smaller supervised models on this training set, we thus avoid bringing in at best irrelevant and at worst misleading information from models trained on a much wider range of data and designed for much more general applications. Having said this, if LLM annotations perform well, it would be straightforward to include them as one of the potential models that we select over, as described in 2.3.

6.3 Alternative NLP approaches

Once open-ended interviews have been conducted and transcribed, our approach of scaling up interpretative qualitative analysis is not the only one researchers could take in analysing the text. In this Section we discuss a range of alternative NLP-based approaches that researchers could take, outlining the advantages of our approach.

6.3.1 Simple word/phrase frequency methods

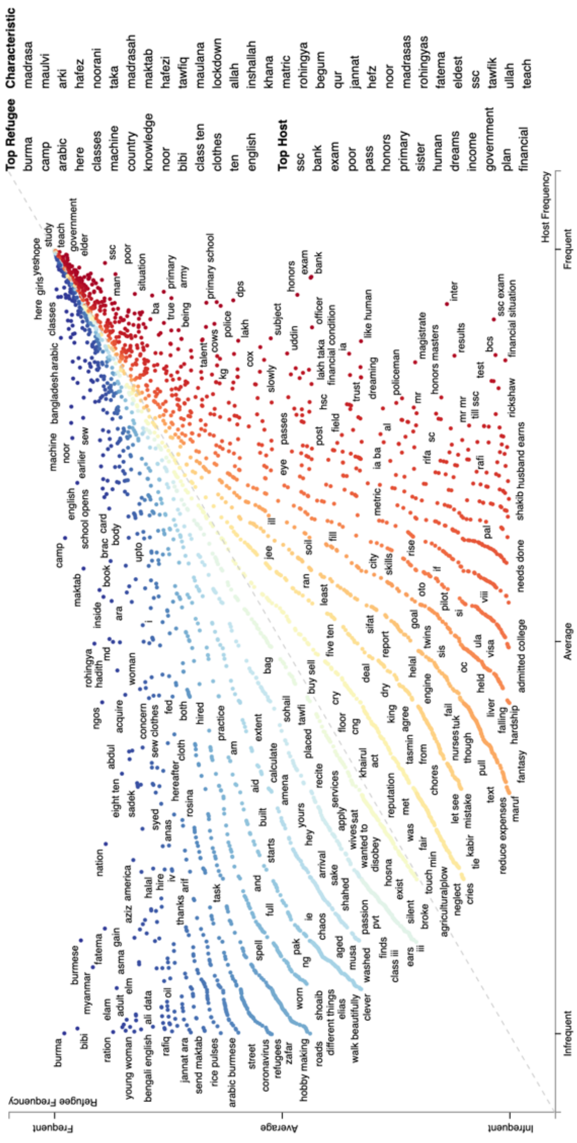
Perhaps the simplest approach to text analysis is to count the occurrence of words and phrases. An example of this can be seen in Figure 9 which shows the frequency of unique terms and bi-grams in interviews with refugees (on the vertical axis) and hosts (on the horizontal axis). Terms that appear frequently for one group but not the other, i.e. terms that appear in the north west and south east of the chart, can be thought of as having a high “keyness” score, in the vocabulary of Ferrario and Stantcheva (2022a). This analysis throws up some interesting patterns. For example, we see far more references to specific education and employment goals for hosts ("ssc exam", "honors masters", "magistrate", "policeman"). However, relating that a these patterns to specific hypotheses around ambition, aspiration and navigational capacity is difficult. Furthermore, the sheer dimensionality of methods based on the raw frequency of all words or phrases across large corpora quickly becomes unmanageable.

On a similar note, we can compare basic summary statistics of the interview transcripts across groups. For example, we find that interviews of parents with a female eldest child are generally slightly longer than those with parents of a male eldest child, while there is no significant difference in length between refugees and hosts. While these sorts of comparisons may be very enlightening in some contexts, in many cases they will not provide strong evidence to a specific research question or hypothesis.

6.3.2 Dictionary methods

Dictionary or lexicon methods are commonly used in social science. These involve counting the number of occurrences of words in pre-specified lists. Many sentiment analysis methods apply this approach by, for example, comparing the frequency of words in a list of "positive" words with the frequency of words in a list of "negative" words. Dictionary methods can be very effective and robust, particularly for analysis of broad concepts like sentiment - VADER (Hutto and Gilbert, 2014) is still regularly used as a benchmark for sentiment analysis methods. Researchers have proposed context-specific dictionaries to measure a wide variety of concepts such as moral values (Enke, 2020), central bank "hawkishness"

Figure 9: N-gram frequency for refugees and hosts



Note: This Figure shows the frequency of words and phrases interviews with hosts on the horizontal axes and refugees on the vertical. Phrases in the north-west thus appear relatively more frequently for refugees and phrases in the south-east appear relatively more frequently for hosts.

(Apel and Grimaldi, 2012) and the tone of financial news Loughran and McDonald (2011). The context-specificity of these word lists is of course a limitation as well as an advantage.[16]

Figure 10: An example of misleading dictionary methods

(a) Prevalence of emotion across groups

Dependent Variable:	Anger	Fear	Trust
Variables			
Refugee status	-0.116	-1.13***	1.00***
	(0.117)	(0.164)	(0.248)
Female eldest child	-0.260**	-0.044	0.290
	(0.108)	(0.151)	(0.228)
Parent's years of education	-0.059***	-0.008	0.065**
	(0.015)	(0.021)	(0.032)

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

(b) Terms most frequently assigned to emotion

Anger	Fear	Trust
money	government	teach
bad	marry	hope
difficulty	god	money
lawyer	bad	school
mastery	difficult	good
treat	pray	doctor
words	problem	happy
broken	medical	marry
fee	difficulty	god
force	die	father

Note: This Figure illustrates how applying dictionary methods from another context can lead to misleading results. Panel (a) shows how the prevalence of terms in different emotion categories, as assigned by Mohammad and Turney (2013), relate to various household characteristics. In panel (b) we see the 10 words in each emotion category that appear most frequently in our transcripts. The font size indicates how often each word occurs.

A major drawback of dictionary methods is thus that they obviously require the existence of a dictionary appropriate for the question and context. As far as we are aware, there are no established dictionaries that measure ambition, aspiration and navigational capacity of parents for their children. Furthermore, even if there were such a dictionary, it is unclear whether applying it in the specific context of Rohingya refugees in Bangladesh. For example the vocabulary that would reveal that a parent in the US has high navigational capacity to achieve their goals for their children would be very different than that for a Rohingya refugee.

[16]In fact, Ashwin *et al.* (2024) suggest that, as tailored dictionaries have been constructed with previous observations in mind, they do not capture unexpected phenomena as well as general purpose methods.

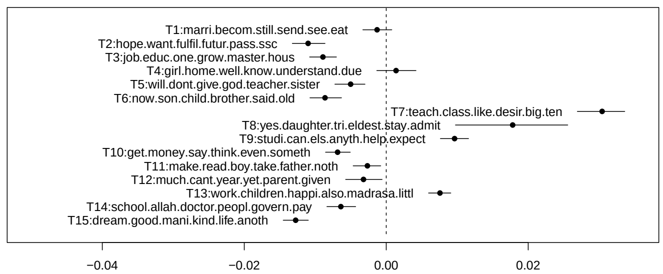
Figure 10 illustrates how dictionary methods might give misleading results in our context. Here, we construct metrics at the interview level based on three emotions from the widely used NRC emotion dictionaries (Mohammad and Turney, 2013): anger, fear and trust. We regress these emotion metrics on the subjects refugee status and whether their eldest child is female, as shown in panel (a) finding some apparently interesting results but misleading results. Parents of female children are apparently less likely to express anger and more likely to express trust, while refugees are less likely to express fear and more likely to express trust. To see that these results are misleading, we can identify the most frequently occurring words in our transcripts that are classified as indicating each emotion by the NRC dictionaries, as shown in panel (b). The presence of these words thus drives the variation in the metric for each of these emotions in our data. It is clear here that many of the terms driving these emotion results are inappropriate in out context. For example, the term "mosque" is classified under anger, "marry" is classified under fear and "money" under trust. In some contexts these classifications may be appropriate, but in interviews with Bangladeshis and Rohingya refugees on their aspirations for their children they are clearly misleading.

6.3.3 Topic modelling

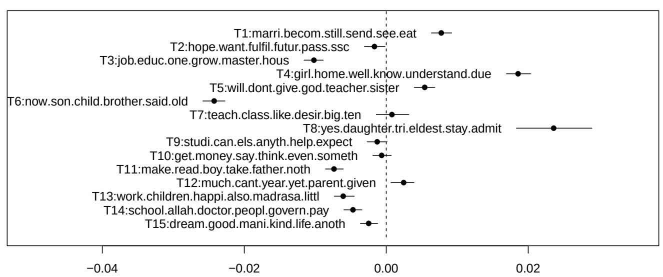
Another popular NLP approach in social science is topic modelling, in which documents are decomposed into latent "topics" that capture the variation across documents. For example, Parthasarathy *et al.* (2019) use a structural topic model (Roberts *et al.*, 2014) to analyse the transcripts of village assemblies in rural India. Topic modelling allows analysis of large corpora without the need for a close reading, but human judgement is necessary to interpret the resulting topics. In some cases, this works well, with the model learning relevant and clearly interpretable topics. However, this is often not the case.

Figure 11: Structural Topic Model results

(a) Refugee effect



(b) Female eldest child effect in topic model



Note: This Figure shows the output of a Structural Topic model estimated on interview transcripts, where the prevalence of each topic is related to household characteristics. The upper panel shows the effect of refugee status on the prevalence of each topic and the lower panel the effect of a female eldest child.

To illustrate this we estimate a 15 topic model using the *stm* R package (Roberts *et al.*, 2019), where each interview is represented as a distribution over topics and the prevalence of these topics are related to our set of household characteristics. We follow standard pre-processing steps, removing stopwords and stemming words. Figure 11 shows the estimated coefficient on the refugee status and female eldest child variables for each topic - a more positive coefficient indicates that this topic is more prevalent in refugee's interviews, conditional on the other household characteristics. Each topic is represented as its five most highly weighted terms. These terms can be used to give an interpretation to each topic.

Giving a clear interpretation to topics can be difficult and, furthermore, imposing the structure of a topic model may obscure subtleties in the text. We can see this from the fact that there are multiple topics which could broadly be interpreted as being about education: topics 3, 7, 9, and 14 all feature terms such as "teach", "educ", "school" and "studi". Some of these topics are more prevalent for refugees and others are more prevalent for hosts, but drawing any firm conclusions from this is difficult. This illustrates how our approach can add value over unsupervised methods.

6.3.4 Semantic similarity

A variety of methods can be used to compare the "semantic similarity" between two documents. Word embedding models such as BERT (Devlin *et al.*, 2018) and RoBerta (Liu *et al.*, 2019) represent individual words or whole documents in numerical vector space. In fact, we include a variety of word embedding models as potential options for the text vectorization in our iQual package (see Appendix C.1). These vector representations of documents can be used as a measure of how closely related documents are, or two cluster documents into groups. In order for patterns in semantic similarity to be interpretable though, we need to move beyond a purely unsupervised approach. A potential way to proceed here is to use word-embeddings to compare the semantic similarity of documents to some paradigm (Acemoglu *et al.*, 2022). This illustrates that moving beyond a purely statistical representation of text to some interpretable metrics requires human input. Our approach can be thought of as a generalisation of approaches that use word embeddings to capture semantic similarity.

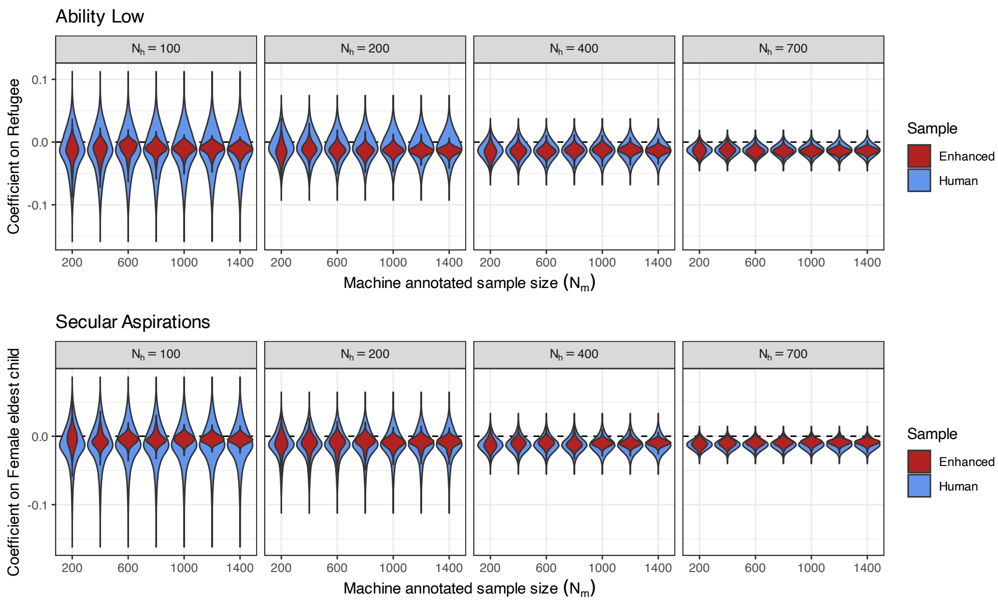
7 How many annotated transcripts are "enough"?

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. We test this by varying the size of both the human annotated (N_h) and machine annotated (N_m) samples. The coding tree is the same across simulations. 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. Our results remain largely the same if at least 400 interviews are annotated. We also show that 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. While these exercises are obviously specific to our context, and may not carry over to other applications, they suggest that in our case annotating more than around four or five hundred transcripts is probably unnecessary.

7.1 Efficiency

The precision of our estimates can be affected by increasing both N_h and N_m - we get stronger results by either annotating more interviews or by conducting additional interviews and machine annotating them. For each model trained on subsamples from 100 to the full 789 human annotated interviews, as described in Section 4.2, we generate predictions for a randomly sampled subset of the unannotated interviews, from 200 to 1,400 (drawing 10 separate samples for each combination of N_h and N_m).

Figure 12: Distribution of regression coefficients of interest with N_h and N_m



Note: This Figure 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) vary. The upper panels represent the coefficient on refugee status for Ability Low (see Figure 8) while the lower panels represent the coefficient on female eldest child for Secular Aspirations (see Figure 7). The distribution of the coefficient estimated on the human sample is in blue and on the enhanced sample is in red. Across panels, from left to right we increase N_h, and 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. The coefficient distributions are calculated through bootstrapping both which interviews are included in the training sample and the coefficient estimate itself.

Our enhanced sample efficiency measure (see Figure 5) increases as extra interviews are added through N_h and N_m. Averaging across all codes, we find adding 100 extra human annotated interviews increases the F statistic in our enhanced sample regressions 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.[17]

We also examine the effect of increasing N_h while holding $N = N_h + N_m$ fixed. Intuitively, we can think of this as human annotating existing interviews, so the overall sample size doesn't increase. Interestingly, while the enhanced sample F statistics 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 - at some point human annotating additional interviews makes little difference.[18]

As an alternative to the very general approach of focusing on the F-statistic across all codes, 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 12 shows how two coefficients of interest from Section 5 (the coefficients on refugee status for Capacity: Ability Low and female eldest child for Aspiration: Secular) 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. Strikingly, the enhanced sample coefficients do not change very much with N_h, again suggesting that enhancing the sample is useful even with very few human annotated interviews.

[17] This efficiency measure for each code as both N_h and N_m vary is shown in Figure D.7.
[18] Results for this exercise are shown in Figure D.8.

7.2 A cost benefit analysis

To give a sense of the trade-off a researcher may face in deciding on how many interviews to conduct and how many to annotate, we conduct a simple cos t benefit analysis exercise. 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 (2021 US dollars) and took around 15 minutes.[19] For given budgets (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: (i) maximising the average F statistic in the enhanced samples for a regression of codes on household characteristics across all codes; (ii) minimising the 95% confidence interval of the refugee status coefficient in the enhanced sample regression on ability low, i.e. upper panels in Figure 12; minimising the 95% confidence interval of the refugee status coefficient in the enhanced sample regression on secular aspiration, i.e. lower panels in Figure 12.

For each objective, increasing either N_h or N_m generally leads to a better outcome, but is more expensive. For a given budget we can thus calculate out the optimal mix of N_m and N_h.[20] Table 5 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. For the most general objective (the average F statistic), our exercise suggests that as the budget increases, it may be preferable to actually only annotated 200 annotations with the expert codes and then use the additional funds for extra machine

[19]The cost for annotation is likely an underestimate as the major difficulty with high quality annotation is finding annotators with the adequate skills. Costs for other studies in other countries and languages may of course vary substantially.

[20]These combinations are shown in Figure D.9

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.

Table 5: Cost Benefit Scenarios

Objective	Budget	N_h	N_m	Price
Average F statistic	**$10,000**	**100**	**600**	$8,700
Ability Low Refugee effect	$10,000	500	200	$9,900
Secular Aspirations Female child effect	$10,000	500	200	$9,900
Average F statistic	**$15,000**	**200**	**1,000**	$15,000
Ability Low Refugee effect	$15,000	500	600	$14,700
Secular Aspirations Female child effect	$15,000	789	200	$14,235
Average F statistic	**$20,000**	**200**	**1,400**	$19,800
Ability Low Refugee effect	$20,000	789	600	$19,035
Secular Aspirations Female child effect	$20,000	400	1,000	$18,000

Note: This Table shoes the optimal N_h and N_m for budgets of $10,000, $15,000 and $20,000 according three objectives: maximising the average F statistic in enhanced sample regressions of code prevalence on household characteristics, and minimising the 95% confidence interval on two specific coefficients. In each case, we find that it is optimal to machine annotate some of the interviews.

This cost benefit analysis exercise is obviously specific to our context, but a similar approach could be used to iteratively determine an appropriate human annotated sample size in other contexts. After annotating some initial minimal number of documents (which would already be necessary in creating the coding tree), we would suggest that researchers test our methodology and, if the results suggest that their sample is underpowered, they can either manually annotate another batch of existing interviews, or conduct some more and machine annotate them.

8 Conclusion

This paper introduces a new method (iQual) to analyse transcripts from interactive open-ended interviews at scale by using supervised natural language processing (NLP) to scale insights from a small set of human-annotated interviews to a larger, representative sample. We rigorously evaluate the reliability and robustness of this method, including guidance on how many human-coded interviews are needed for researchers with budget constraints. Our approach outperforms alternatives such as unsupervised NLP techniques and direct annotation using large language models. We apply it to 2,200 open-ended interviews on parental aspirations for children, conducted with Rohingya refugees and their Bangladeshi hosts. Our analysis shows that open-ended interviews could enrich the economic study of aspirations by incorporating philosophical and anthropological concepts—such as moral and religious values, and the capacity to pursue them. This method also uncovers novel findings, including a potential form of migrant selection: while Rohingya refugees have lower education levels than their Bangladeshi hosts, they may be positively selected on navigational capacity.

Interpretative qualitative analysis, which is a common tool in anthropology, sociology and related disciplines but not used by economists, 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 analysed 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 structure 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. It could also be potentially of 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. Our method provides a feasible option for researchers who wish to take advantage of the benefits of interpretative qualitative analysis without sacrificing the benefits of large samples.

Affiliations

[1]Maastricht University

[2]World Bank Group

[3]Qual Analytics

References

Acemoglu, D., Mühlbach, N.S. and Scott, A.J. (2022). 'The rise of age-friendly jobs', *The Journal of the Economics of Ageing*, vol. 23, p. 100416.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019). 'Optuna: A next-generation hyperparameter optimization framework', .

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*, vol. 672(1), pp. 143–161.

Andre, P., Haaland, I., Roth, C., Wiederholt, M. and Wohlfart, J. (2024). '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*, vol. 7(2).

Ashwin, J., Chhabra, A. and Rao, V. (2025). 'Using large language models for qualitative analysis can introduce serious bias', *Sociological Methods & Research*.

Ashwin, J., Kalamara, E. and Saiz, L. (2024). 'Nowcasting euro area gdp with news sentiment: a tale of two crises', *Journal of Applied Econometrics*, vol. 39(5), pp. 887–905.

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

Beaman, L., Duflo, E., Pande, R. and Topalova, P. (2012). 'Female leadership raises aspirations and educational attainment for girls: A policy experiment in india', *Science*, vol. 335, pp. 582–586.

Berezkin, Y.E. (2015). 'Folklore and mythology catalogue: its lay-out and potential for research', *The Retrospective Methods Network*, vol. 1(S10), pp. 58–70.

Bernard, T. and Taffesse, A.S. (2014). 'Aspirations: An approach to measurement with validation using ethiopian data', *Journal of African Economies*, vol. 23, pp. 189–224.

Birgier, D.P., Lundh, C., Haberfeld, Y. and Elldér, E. (2016). 'Self-selection and host country context in the economic assimilation of political refugees in the united states, sweden, and israel', *International Migration Review*.

Black, S.E. and Devereux, P.J. (2011). 'Recent developments in intergenerational mobility', *Handbook of labor economics*, vol. 4, pp. 1487–1541.

Black, S.E., Devereux, P.J. and Salvanes, K.G. (2005). 'Why the apple doesn't fall far: Understanding intergenerational transmission of human capital', *American economic review*, vol. 95(1), pp. 437–449.

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

Borjas, G.J., Kauppinen, I. and Poutvaara, P. (2019). 'Self-selection of emigrants: Theory and evidence on stochastic dominance in observable and unobservable characteristics', *The Economic Journal*, vol. 129(617), pp. 143–171.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. *et al.* (2020). 'Language models are few-shot learners', *Advances in neural information processing systems*, vol. 33, pp. 1877–1901.

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)*, vol. 8(2), pp. 1–20.

Chiquiar, D. and Hanson, G.H. (2005). 'International migration, self-selection, and the distribution of wages: Evidence from mexico and the united states', *Journal of political Economy*, vol. 113(2), pp. 239–281.

Chiswick, B.R. (1999). 'Are immigrants favorably self-selected?', *American Economic Review*, vol. 89(2), pp. 181–185.

Chopra, F. and Haaland, I. (2023). 'Conducting qualitative interviews with ai', CESifo Working Paper.

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*, vol. 47(1), pp. 1–2.

Detering, N.M. and Waters, M. (2018). 'Flexible coding of in-depth interviews: A twenty-first century approach', *Sociological Methods and Research*, vol. 50(2)), pp. 708–738.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*.

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

Enke, B. (2020). 'Moral values and voting', *Journal of Political Economy*, vol. 128(10), pp. 3679–3729.

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

Favara, M. (2017). 'Do dreams come true? aspirations and educational attainments of ethiopian boys and girls', *Journal of African Economics*, vol. 26, pp. 561–583.

Ferrario, B. and Stantcheva, S. (2022a). 'Eliciting people's first-order concerns: text analysis of open-ended survey questions', .

Ferrario, B. and Stantcheva, S. (2022b). 'Eliciting people's first-order concerns: Text analysis of open-ended survey questions', *AEA Papers and Proceedings*, vol. 112, pp. 163–169.

Filippini, M., Leippold, M. and Wekhof, T. (2021). 'Sustainable finance literacy and the determinants of sustainable investing', *Swiss Finance Institute Research Paper*, vol. 1(22-02).

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*, vol. 38(1), pp. 115–132.

Fruttero, A., Muller, N. and Calvo-Gonzalez, O. (2025). 'The power and roots of aspirations', *World Bank Research Observer*, vol. 50, pp. 58–103.

Geiecke, F. and Jaravel, X. (2024). 'Conversations at scale: Robust ai-led interviews with a simple open-source platform', *Available at SSRN 4974382*.

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

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

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

Gilardi, F., Alizadeh, M. and Kubli, M. (2023). 'Chatgpt outperforms crowd-workers for text-annotation tasks', *Proceedings of the National Academy of Sciences*, vol. 120(30), p. e2305016120.

Haaland, I.K., Roth, C., Stantcheva, S. and Wohlfart, J. (2024). 'Measuring what is top of mind', National Bureau of Economic Research.

Houde, S. and Wekhof, T. (2021). 'The narrative of the energy efficiency gap', *Economics Working Paper Series*, vol. 21.

Hutto, C. and Gilbert, E. (2014). 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', .

Jayachandran, S., Biradavolu, M. and Cooper, J. (2021). 'Using machine learning and qualitative interviews to design a five-question women's agency index', 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*, vol. 1, pp. 33–43.

Liew, J.S.Y., McCracken, N., Zhou, S. and Crowston, K. (2014). 'Optimizing features in active machine learning for complex qualitative content analysis', .

Liu, A. and Sun, M. (2023). 'From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews', *arXiv preprint arXiv:2312.01202*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692*.

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

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

Mohammad, S.M. and Turney, P.D. (2013). 'Nrc emotion lexicon', National Research Council Canada.

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*, vol. 113(3), pp. 623–640.

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

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

Roberts, M.E., Stewart, B.M. and Tingley, D. (2019). 'Stm: An r package for structural topic models', *Journal of Statistical Software*, vol. 91, pp. 1–40.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G. (2014). 'Structural topic models for open-ended survey responses', *American journal of political science*, vol. 58(4), pp. 1064–1082.

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

Roy, S., Morton, M. and Bhattacharya, S. (2018). 'Hidden human capital: Self-efficacy, aspirations and achievements of adolescent and young women in india', *World Development*, vol. 111, pp. 161–181.

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

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

Stantcheva, S. (2020). 'Understanding economic policies: What do people know and learn?', *Unpublished Manuscript, Harvard University*, pp. 2309–2369.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. (2023). 'Llama: Open and efficient foundation language models', .

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D. *et al.* (2022). 'Chain-of-thought prompting elicits reasoning in large language models', *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837.

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

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', .