Policy Research Working Paper 10255

Fewer Questions, More Answers

Truncated Early Stopping for Proxy Means Testing

Tim Ohlenburg
Juul Pinxten
Daniel Fricke
Fabio Caccioli

WORLD BANK GROUP
Poverty and Equity Global Practice
December 2022

Policy Research Working Paper 10255

Abstract

The assignment of social programmes to their target population, known as targeting, is key to effective policy implementation. Proxy means testing is a widely used targeting approach where means testing is infeasible due to economic informality. This paper proposes a novel, practically feasible assessment approach that aims to reduce average proxy means test data collection costs, or allow more extensive data collection within a given resource envelope. Combining variable selection and prediction intervals, it develops a household-level truncated early stopping algorithm, which can reduce average interview length while maintaining predictive accuracy close to a standard proxy means test baseline. Applying the approach to Indonesian data, simulation of a 40 percent population coverage programme shows that targeting questionnaires could be shortened by 61 percent while maintaining PMT-level accuracy. A case study of a large health insurance programme in an urban area suggests that the share of intended beneficiaries who are among the targeted population can potentially be increased from 65.6 percent to 78.3 percent if enumerators conducted more of the shorter surveys that the truncated early stopping algorithm generates.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at email@timohlenburg.com, jpinxten@worldbank.org, d.fricke@ucl.ac.uk, and f.caccioli@ucl.ac.uk.

Produced by the Research Support Team

Fewer Questions, More Answers: Truncated Early Stopping for Proxy Means Testing

Tim Ohlenburg*[1], Juul Pinxten[2], Daniel Fricke[1], and Fabio Caccioli[1]

[1]University College London (UCL)
[2]World Bank

JEL I32, C83, O22
Keywords proxy means testing, adaptive survey design, policy simulation

*This paper has benefitted from the patient guidance and comments of Utz Pape, Virgi Agita Sari, Gracia Hadiwijaja, and Sailesh Tiwari. We are also grateful to reviewers Takaaki Masaki, Matthew Wai-Poi, and Phillippe Leite, whose constructive feedback improved contents and presentation.

1 Introduction

Low- and middle-income countries spend an average of 1.5% of GDP on social assistance [World Bank, 2022]. These public programmes aim to protect the poor and vulnerable from shocks, help them cope with crises, support investment in human capital and enjoy a decent standard of living throughout the lifecycle [ILO, 2021]. Policymakers design a social protection programme with a specific target population in mind. A crucial aspect of policy implementation, commonly referred to as the targeting problem, is how to identify households or individuals who are members of the target population and therefore eligible for the programme. Due to limited fiscal resources, social assistance tends to be limited to the most economically disadvantaged segment of society. Consequently, socio-economic criteria are a prevalent eligibility category [Grosh et al., 2022], whether applied in isolation – such as in a food subsidy programme for the poor – or in combination with another eligibility category, such as conditional cash transfer programmes that support poor families with school-age children.

In settings where nearly all households' income is reported to the tax authority, means testing via income tax records can determine socio-economic eligibility. However, countries' average informal employment share is above 50% in Asia Pacific, South Asia and Sub-Saharan Africa, while the poverty rate of the informally employed is around six times higher than that of those in formal employment [Ohnsorge and Yu, 2021]. This paper is concerned with such settings, where means testing is unviable due to a lack of tax records among poorer households. There, proxy means testing (PMT) has become a widespread targeting approach since its description by Grosh and Baker [1995].

PMT is a statistical approach that relies on two large-scale surveys. The first, conducted at regular intervals in most countries, is a socio-economic population sample survey that serves as training data for the PMT predictive model. The second is a targeting survey that collects current data to determine household eligibility. Due to the large population coverage of many social protection programmes, targeting surveys typically need to be administered to a significant share of the population[1]. Given the high cost of such extensive, country-wide data collection efforts, targeting surveys are typically conducted on a multiannual basis when carried out as a national survey sweep [Lindert et al., 2020]. Infrequent surveys delay assessment of new programme applicants and re-assessment of existing beneficiaries, resulting in programmes' irresponsiveness to changing household circumstances.

If budgets were not limited, the most accurate PMT approach would be

[1] For example, the social registries that store targeting data covered 87%, 75% and 40% of Pakistan's, Brazil's and Indonesia's populations respectively in 2015-17 [Leite et al., 2017], amounting to total records of ca. 360 million people.

to assess all households, census-like, and to apply an extensive questionnaire that asks all potentially relevant questions. In practice, administrative budgets of social protection programmes are limited, so that neither all households will be assessed nor all questions asked. When eligible households are missed from the assessment process, exclusion errors will arise no matter how accurate the PMT model. For Indonesia, World Bank [2012] identifies insufficiently broad assessment as a key area for improving targeting outcomes, and its simulation shows that a full survey would lower exclusion error significantly. Since budget constraints normally preclude universal assessment, a practical option may be to streamline the data collection process by asking fewer questions. If this could be done without significantly compromising model accuracy, more households could be assessed for the same budget, which would improve targeting outcomes at no additional cost.

The aim of this paper is to develop such a method, achieving a similar level of targeting accuracy as a standard PMT while asking fewer questions. We propose an alternative to the PMT algorithm that applies a measure of predictive uncertainty to household consumption estimation. We use prediction intervals of a quantile regression-style estimator to identify those households for which consumption is most likely above or below the socioeconomic eligibility threshold. By adjusting the intervals as data is collected, enumeration can be stopped early (in the sense that not all questions need to be asked) for a substantial share of households. In addition, questionnaire length is limited to the most predictive subset of questions. Our approach combines these mechanisms to expose a trade-off between predictive accuracy and questionnaire length that policymakers can set according to their preferences or constraints. In contrast to more computationally demanding methods proposed for an adaptive poverty classification, its limited computational requirements make deployment feasible on the modest mobile devices commonly used for digital data collection in low- and middle income countries. The financial resources freed up by this approach could thus be directed to more regular or more extensive targeting surveys.

In our empirical application for Indonesia, we find that the reduction in questionnaire length compared with the PMT is rather substantial. For a 40% coverage programme, the number of questions can be reduced by 61% (from an average of 22.8 to 8.9) when maintaining the exclusion error rate at the PMT baseline of 26.44%. Tentative estimates for an urban setting, namely Jakarta, in which we simulate a health insurance programme with 50% population coverage, suggest an increase in the proportion of eligible recipients from 65.6% under PMT-based assignment to 78.3% under the proposed method, assuming enumerators can survey additional households by deploying shorter, adaptive questionnaires.

The paper is structured as follows. Section 2 reviews the relevant literature, outlines the PMT process, and provides context for the data used in subsequent simulations. Section 3 describes the components and functioning of our novel targeting method. The results in Section 4 compare the performance of our algorithm against a policy baseline. The Jakarta case study in Section 5 presents an approach to maximize the inclusion of intended beneficiaries by balancing survey coverage and accuracy in an urban setting. Finally, Section 6 discusses the results and their implications.

2 Background

This section describes the standard approach to PMT as the foundation of our proposed method. It sets the scene by linking the research to the literature, and by framing PMT as a predictive modelling task with specific design considerations. It then describes the PMT algorithm and provides some context on its use in Indonesia, including the data collected for its implementation.

2.1 Literature

This paper proposes a novel algorithmic approach to targeting and its general context lies within the literature on targeting social programmes, and particularly the PMT approach. The computational method and emphasis on predictive modelling link it to the growing literature on machine learning methods in targeting. The final related area we touch on is the literature on adaptive designs for survey cost reduction.

The seminal paper on PMT as a targeting mechanism is the work of Grosh and Baker [1995], who described a statistical welfare estimation approach that had been developed in Chile in the 1980s. With PMT's increasing popularity, Coady et al. [2004] conducted a systematic cross-country assessment that suggested it to be the best option where means testing is infeasible, but that its effectiveness is often limited and very context-specific. More recently, Brown et al. [2018] found that PMT could stand for Poor Means Test, as targeting errors are often so high that untargeted benefit programmes or simpler scorecard methods work nearly as well across their country examples.

Indonesia has long been a productive place for targeting research, supported by progressive policymakers, their relationships with researchers and the international development community. An influential output of this collaboration was Alatas et al. [2012], who showed in a randomized controlled trial setting that PMT results in somewhat better targeting outcomes than community-based targeting, but only when a consumption measure rather than community perceptions of poverty were used as eligibility yardstick. Relevant to survey design is an experiment by Banerjee et al. [2020], which suggested that the inclusion of certain assets in a targeting survey does not distort Indonesians' buying behaviour, but that there is evidence of strategic misreporting. Similarly, Camacho and Conover [2011] in Latin America and Niehaus et al. [2013] in South Asia find evidence of manipulation on the survey administration side. Tohari et al. [2019] made a case for considering the full set of programmes with socio-economic criteria when conducting targeting simulations via their extensive linking of Indonesian survey data.

A growing literature considers PMT from a predictive modelling perspective. McBride and Nichols (2018) highlighted the importance of using out-of-sample validation data in model selection, which has been adopted in Indonesia. A simplified scorecard approach supported by machine learning prediction was suggested by Kshirsagar et al. [2017], who showed impressive results with very limited data collection needs for Zambian data. A systematic evaluation of machine learning methods for construction of the predictive model in PMT was carried out by Areias and Wai-Poi [2022] with data from 12 African countries, but it found that accuracy gains tend to be limited and context specific; no clear machine learning works best across the board. A similar conclusion emerged from a study of Indonesia's PMT model [Ohlenburg, 2020].

The model presented here has a household-specific stopping criterion as its adaptive component. Adaptive methods link to extensive literatures on the design of experiments, adaptive algorithms and active learning. As an example of the latter, Saar-Tsechansky et al. [2009] studied active feature acquisition when data collection is costly and variables differ in their information value. They proposed a framework that aims to maximize the expected utility of each data item and tailors the sequence and length of data acquisition to each instance, achieving a cost reduction for a given level of accuracy.

A line of research that focuses on the aspect of fairness in targeting emerged in the work of Noriega-Campero et al. [2020]. It puts an emphasis on achieving equitable accuracy rates across subgroups of the population, thereby avoiding systematic disadvantages in programme eligibility for specific groups. In contrast to the evidence of limited accuracy improvements mentioned in the machine learning for targeting literature mentioned above, this work also suggests scope for meaningful improvements in targeting accuracy from changes in data preparation, such as the use of a deep feature embedding, and predictive modelling methods including neural networks.

Bakker et al. [2019] pursued the fairness theme further in a paper that leverages reinforcement learning for selection of fair targeting questions that tailor the questionnaire sequence to each household according to group membership and other characteristics. Closely related to this paper is Bakker et al. [2021], an adaptive design that computes a household-specific sequence of questions and stops when a predictive certainty threshold is met in classifying households as poor or non-poor. Our approach differs in the use of a uniform question sequence and the prediction of a continuous consumption level. Estimation of the household consumption level enables us to rank households, which is important for adjusting eligibility to meet coverage targets and for the common case of multi-programme assessment highlighted by Tohari et al. [2019].

In view of the huge logistical scale of many targeting surveys, sample population surveys are perhaps the most closely related information collection exercise. Looking at their economic aspects, Groves and Heeringa [2006] proposed a responsive survey design to reduce survey cost while maintaining accuracy. Focusing on settings with cost and operational constraints on data collection, Pape and Mistiaen [2018] proposed an effective imputation approach to estimate the population distribution of consumption rather than a full LSMS-style survey. A similar focus on a reduced length survey is Christiaensen et al. [2021], who investigated the use of components of consumption to estimate household values but find both theoretical and empirical problems with this approach.

2.2 PMT design

Three essential design criteria, echoed in Grosh et al. [2022], shape PMT:

- *Accuracy.* Identifying the intended beneficiaries accurately is the essential paradigm of a targeting mechanism.
- *Cost.* The multiannual nation-wide PMT survey sweeps common in low- and lower-middle income countries require major fiscal outlays. The resource needs of on-demand registration, which countries with sufficient administrative capacity tend to invest in, are also high and imply the need for concise surveys that economize enumerator time.
- *Verifiability.* Survey responses determine PMT outcomes, creating a monetary incentive for misreporting personal and household characteristics towards responses that are associated with poverty[2]. As a result, targeting surveys are often restricted to variables that are observable by enumerators, such as physical assets, or that are verifiable via documentary evidence.

In comparison to other commonly used targeting methods such as geographic targeting, community-based targeting, means testing or hybrid means testing, PMT uses easy-to-verify characteristics. It is most appropriate when informality is high and when some form of household specific ranking is desired. At lower levels of informality in an economy, means or hybrid means tests would be more appropriate and yield higher accuracy than a PMT [Grosh et al., 2022].

[2] Banerjee et al. [2020] provide evidence of such strategic behaviour, which creates an unfair advantage for those willing to be dishonest.

2.2.1 PMT process

To understand how a PMT aims to achieve these design considerations, we break the approach down into its components. A PMT is a combination of survey and statistical modelling work. It can be described as illustrated in fig. 1:

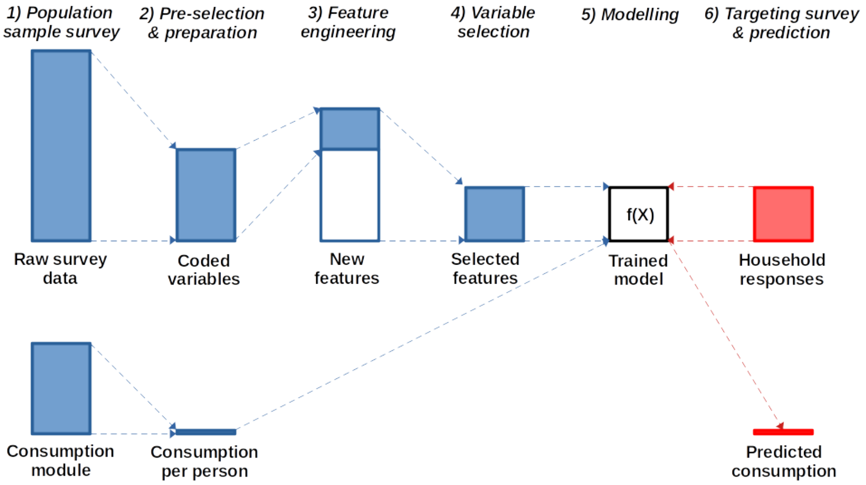


Figure 1: PMT modelling stages

1. *Population sample survey.* The country's statistics authority conducts a nationally representative survey to monitor socio-economic conditions, independently of the PMT. This survey has two elements from a PMT perspective: (a) a consumption module that is aggregated to the dependent variable, and (b) the remainder of the survey contains the set of potential independent variables.

2. *Pre-selection & preparation.* In variable pre-selection, a subset of the non-consumption variables is chosen. As per the design criteria, it should be both verifiable and also avoid perverse incentives for behavioural changes that may ultimately be harmful, such as "do your children attend school regularly?"

3. *Feature engineering.* The creation of additional features, such as variable transformations, interaction terms, group markers etc. is an important aspect of predictive modelling in data-constrained settings.

4. *Variable selection.* The combined set of potential features that emerges from step 3 is usually large, which would result in high survey costs if all were included in the targeting survey. The task of variable selection

is to identify those features – denoted as X – that offer most predictive power. Some approaches, such as stepwise regression, perform the selection simultaneously with model building, but the two tasks can also be split.

5. *Modelling.* The modelling step involves the development of a supervised learning model $f(X)$ that predicts consumption per household member. Below we expand on the PMT approach to variable selection and modelling, and then propose an adjusted algorithm.

6. *Targeting survey & prediction.* Operationalizing a PMT requires the set of variables X to be queried from a list of potential beneficiaries. In addition to the features, this survey may include household characteristics that determine categorical programme eligibility, such as the number of school-age children, verification of household characteristics or identity via inspection of administrative documents. The model $f(X)$ is applied to produce consumption estimates for the set of potential beneficiaries. Combining categorical and socio-economic eligibility, the list of beneficiaries is determined either by a ranking of household values (in case of a beneficiary quota) or via an absolute threshold that assigns eligibility if a given household's income falls below it.

To consider the link between these steps and the design considerations outlined before, note that both cost and verifiability are influenced by the questionnaire design of the targeting survey. The number and detail of questions and the associated verification routines are an important cost lever, especially in urban areas where travel times between households are short for enumerators. In terms of the PMT process shown in fig. 1, verifiability is considered in the pre-selection step. The variable selection aspect of the modelling step also influences costs by shrinking the volume of questions. The implicit policy objective of the PMT algorithm is to achieve the best possible accuracy for a set of pre-selected questions. As such, the emphasis of the standard PMT approach is accuracy, rather than on cost reduction.

The focus of this paper is on modelling, but note that targeting survey design and eligibility determination are also impacted by modelling choices: the targeting survey consists primarily of variables selected, and the predictive model is used to score eligibility. The key statistical challenge in PMT design is model selection, particularly to limit the number of independent variables without sacrificing accuracy. The number of possible variable combinations is enormous in most settings due to the wide extent of the sample population survey. It is computationally infeasible to try out any but a fraction of these combinations within a chosen predictive model, and PMT manages this challenge with the following algorithm.

2.3 PMT algorithm

PMT uses a 'greedy' approach to break the intractable variable selection problem down into a sequence of manageable tasks. At each step, it selects the best currently available option, until a stopping criterion is triggered. Such a myopic procedure can result in a selection that is very different to the optimal subset. However, the lack of theoretical guarantees is outweighed by its feasibility and results that tend to be credible, as evidenced by widespread policy adoption. A further benefit from the perspective of this paper is that a greedy algorithm produces a sequence of nested variable sets, which is a requirement for the adaptive approach laid out below. The original PMT algorithm, implicit in Grosh and Baker [1995], uses stepwise regression as the basic algorithm. It consists of the following key components:

- *Predictive model.* Ordinary least squares (OLS) regression is the workhorse of PMT. Starting with an intercept, one OLS model is trained for each un-queried variable.
- *Selection criterion.* The convention in stepwise regression tends to be the use of the Bayesian or Akaike information criterion for in-sample model selection [Friedman et al., 2001]. In forward selection, each variable is considered in turn, and that which provides the largest increase in the information criterion is added to the current variable set.
- *Stopping criterion.* Variable selection continues until the best available model in the current step offers no improvement in the chosen information criterion.

The PMT algorithm is described algorithm 1 below, where X is a design matrix with partitions, X_{sel}, X_{tmp}, and X_{cand} are column-wise partitions thereof, y is the consumption level per household member that PMT estimates to determine eligibility, $f()$ is an OLS predictive model, AIC is the Akaike information criterion [Akaike, 1998].

2.4 Data and background

The Indonesia context

Indonesia uses a social registry to implement its PMT for targeting multiple programmes. Much of the country's population lives clustered above the poverty line and despite marked improvements in welfare, vulnerability remains substantial. In 2018, the poverty rate was 9.8%, but 28% of Indonesians lived below 1.5 times the poverty line. As a result, a relatively small income shock can push around 20% of Indonesian households into poverty. Those living above this vulnerability line but below 3.5 times poverty line

Algorithm 1 Standard PMT (forward selection)

```
Input: Training data $(\mathbf{X}, \mathbf{y})$
  $\mathbf{X}_{sel} \leftarrow \mathbf{1}$
  $f() \leftarrow \texttt{fit}(\mathbf{X}_{sel}, \mathbf{y})$                         ▷ Predictive model
  $\hat{\mathbf{y}}_{sel} \leftarrow f(\mathbf{X}_{sel})$
  $(\mathbf{X}_{tmp}, \hat{\mathbf{y}}_{tmp}) \leftarrow (\mathbf{X}_{sel}, \hat{\mathbf{y}}_{sel})$
  while $\exists \mathbf{X}_{:,j} \in \mathbf{X}, \notin \mathbf{X}_{sel}$ do:
      for each $\mathbf{X}_{:,j} \in \mathbf{X}, \notin \mathbf{X}_{sel}$ do:
          $\mathbf{X}_{cand} \leftarrow \texttt{append}(\mathbf{X}_{sel}, \mathbf{X}_{:,j})$
          $f() \leftarrow \texttt{fit}(\mathbf{X}_{cand}, \mathbf{y})$
          $\hat{\mathbf{y}}_{cand} \leftarrow f(\mathbf{X}_{cand})$
          if $AIC(\mathbf{y}, \hat{\mathbf{y}}_{cand}) < AIC(\mathbf{y}, \hat{\mathbf{y}}_{tmp})$ then:      ▷ Selection criterion
              $(\mathbf{X}_{tmp}, \hat{\mathbf{y}}_{tmp}) \leftarrow (\mathbf{X}_{cand}, \hat{\mathbf{y}}_{cand})$
          end if
      end for
      if $AIC(\mathbf{y}, \hat{\mathbf{y}}_{tmp}) >= AIC(\mathbf{y}, \hat{\mathbf{y}}_{sel})$ then:      ▷ Stopping criterion
          break
      end if
      $(\mathbf{X}_{sel}, \hat{\mathbf{y}}_{sel}) \leftarrow (\mathbf{X}_{tmp}, \hat{\mathbf{y}}_{tmp})$
  end while
```

comprise 46%, indicating that about 75% of the population live on incomes below middle- and upper-class levels. The consumption distribution in fig. 2 shows how tightly clustered the consumption level of a large part of the population is, in turn highlighting the challenge of targeting social assistance programs to the poorest.

The country's PMT was revised in 2015 to consist of 514 distinct district-level models, to account for the differences within a nation that spans geographically and economically diverse regions. Since the introduction of the Unified Database for Targeting (now referred to as DTKS), improved targeting outcomes were seen for social protection programmes that adopted it [World Bank, 2017]. The country's first unconditional cash transfer in 2005 covered more than a third of the population with the aim to protect them from a reduction in the fuel subsidy. Beneficiary incidence – the share of total beneficiaries found in a welfare grouping – in the poorest 20% of the population comprised 36%. When an unconditional cash transfer was launched for a third time in 2013, its coverage was about 40 percent, though beneficiary incidence increased to 40%. Similarly, a conditional cash transfer for families launched in 2007 saw an increase in beneficiary incidence in the poorest 20% from 39% to 44% between 2010 and 2018[3]. Its targeting

[3] Based on Holmemo et al. [2020] and World Bank staff calculations from SUSENAS 2010/2018.

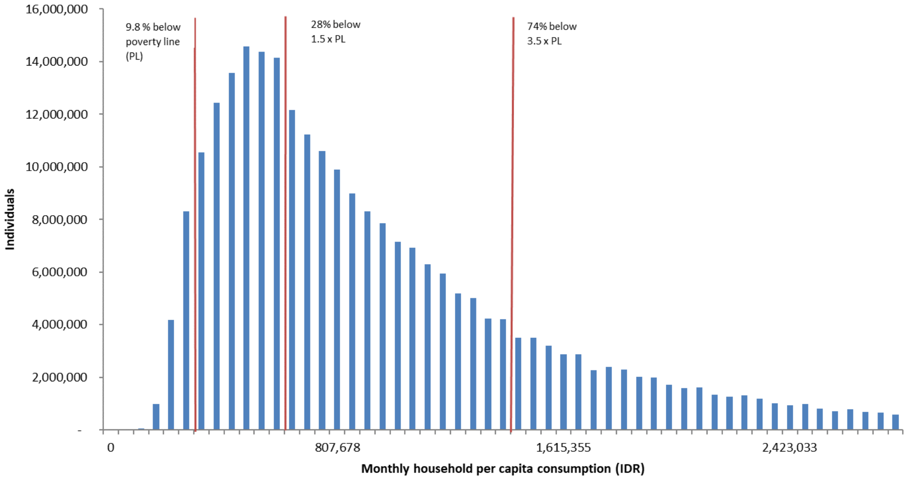


Figure 2: Indonesia's distribution of household consumption per capita, 2018

Source: Holmemo et al. [2020], World Bank staff calculations from SUSENAS 2018

outcomes are on par with similar programs around the world, though exclusion errors remain a major issue. Whilst the programme allocates over 71% of total program benefits to the poorest 40%, about one third of eligible households in the poorest decile still do not receive the program due to being excluded from consideration or being misclassified as non-poor.

Survey questions

The primary data source for Indonesia's PMT and for this paper is the National Socio-Economic Survey (SUSENAS). Each year of this bi-annual population sample survey contains responses from around 300,000 individuals and is representative of the country's 514 districts. Stratified by year and district, the survey rounds for the five years of 2015-19 are split 80-20 into training and test sets of circa 1.2 and 0.3 million observations respectively. SUSENAS includes an extensive consumption module and a living conditions module from which the targeting survey questions are drawn. In the modelling, the consumption data are adjusted for prices and expressed as a logarithm per household member. The targeting survey mirrors the DTKS social registry administered by the Ministry of Social Affairs. Table 1 displays the variables that comprise the feature set.

Table 1: Overview of Indonesian targeting survey variables

Type	Theme	Variables included
Core	Demographics	Household size & HH size^2, number of people by age groups by gender, urban/rural, family structure, family smart card
Optional	Housing	Materials used in the floor, wall, roof, source of drinking water and cooking water, type of lighting & cooking fuel, toilet facilities, septic tank, floor space per capita, ownership status.
	Assets	Household ownership of: motorcycle, car, computer, fridge, boat, motorboat, phone, water heated, air conditioning
	Education	Household members' total levels of educational attainment and enrolment
	Employment	Employment status, employment sectors

Baseline PMT

The Indonesian PMT implemented in the country's social registry provides the baseline method for the early stopping algorithm's performance. Its construction uses the data described above, and follows the canonical PMT approach in Algorithm 1 except for two significant changes. The first is in line with McBride and Nichols [2018], who pointed out that PMT is an out-of-sample prediction task and that a cross-validation approach to variable selection leads to better performance than in-sample measures such as information criteria. Accordingly, we substitute selection via an information criterion with 5-fold cross-validation and a mean squared error (MSE) selection criterion. The stopping criterion is triggered when the cross-validated error no longer declines. MSE is the preferred metric as it aligns with the objective function of predictive model OLS[4]. The second change is to start estimation from a set of core questions required for administrative reasons or known up front. The rationale is that the algorithm should leverage both mandatory and pre-existing data to achieve a higher starting accuracy, rather than to ignore available information and start estimation from scratch. Table 1 displays the assignment of the two variable types.

A separate PMT model is built for each of the 514 districts. Although our baseline follows the current policy practice, there are also minor differences

[4] Areias and Wai-Poi [2022] show that optimizing for MSE does not necessarily minimize EER, which is coverage-level dependent, but it provides a consistent objective for multi-programme targeting across coverage levels

in its construction in terms of the data, and it should not be understood as equivalent to the official PMT. The adjustments include use of different years of SUSENAS to the current set, corresponding to data available when this project was initiated, as well as the simplification of the dataset. We eliminate a number of household demographic and work status interaction terms that were too numerous for the sequential approach proposed below, while offering only a marginal predictive gain. Despite these differences, the targeting metrics are of broadly comparable magnitude[5], and we would not expect a change in the qualitative conclusions if the current PMT were used instead.

3 Methods

We already identified accuracy, cost and verifiability as the three major design considerations of a statistical targeting method. An additional consideration for methods that require on-the-fly prediction in the resource-constrained, often geographically remote setting of PMT deployment is that they need to be computationally feasible. The approaches proposed here meet this criterion by virtue of being relatively simple to implement and deploy on hardware with modest storage, memory and computational power at inference time. If we consider verifiability to be embedded in the selection of variables that can be queried reliably, accuracy and cost remain as desirable outcomes. A more extensive variable set improves the predictive power of household data, as long as the additional variables contain additional information about its consumption level, but it also requires higher collection cost. Consequently, statistical targeting requires a choice along the length-accuracy trade-off in a resource-constrained setting.

The standard PMT approach described above has become widespread since its publication by Grosh and Baker [1995] to design models that use a moderate number of explanatory variables that also achieve a suitable level of targeting accuracy. In its pure form of automated variable selection, it selects the point on the accuracy-vs-enumeration-cost spectrum that maximizes model accuracy. This section revolves around the idea that other points – and especially earlier points – along the variable sequence should also be considered. A shorter questionnaire may be less accurate, yet it may result in lower overall exclusion due to the consideration of more households. We suggest methods which replace the single point with a range of options that allow policymakers to select a length-accuracy profile that is suitable for their conditions.

[5] Klasen et al. [2016] estimated exclusion errors for a forward stepwise regression model of 27% for a simulated 50% program coverage. The forward stepwise model constructed mirrors closely the approach taken in generating PMT rankings in the update of the UDB/DTKS in 2015. Appendix A shows a PMT exclusion error rate of 21% at 50% coverage.

This section introduces three approaches. The first, truncated questionnaires, is a family of methods that includes the standard PMT, and which varies questionnaire length uniformly for all respondents. The second, early stopping, employs a measure of predictive uncertainty as a household-level stopping criterion. Third, Truncated Early Stopping (TrESt), combines the other two methods to leverage their respective strengths. Before an explanation of the methods, the first subsection describes two modelling preliminaries.

Modelling preliminaries

Predictive model In line with standard econometric practice, OLS regression has been the predictive model of choice in the PMT literature. In contrast, this paper uses a gradient boosting machine [Friedman, 2001] as it provides greater predictive accuracy, deals efficiently with large datasets, offers variable interactions without the need for explicit interaction terms, and provides a flexible and coherent framework for all computational tools deployed below except the group lasso. Although machine learning models are often considered less transparent than linear ones, the interpretation of feature importance for tree-based models – described as a variable selection approach below – and the option to use explanation tools such as SHAP [Lundberg and Lee, 2017] that is grounded in Shapley values provide a similar degree of transparency, especially when considering that a causal interpretation of model parameters is unsuitable in a predictive modelling context (see Shmueli [2010]).

Grouped variables Given our emphasis on enumeration cost, the grouping of variables needs to be considered in more detail than in a standard setting. The first of two related aspects concerns derived features that are generated from other variables, such as squared terms, dummy/ one-hot encodings. Although they may enter the predictive model as a stand-alone feature, the underlying information is queried through a question for another variable. Such items should be treated as a single feature, both from a questionnaire length and a variable selection perspective. In this vein, we adjust the various methods below to treat grouped variables as single items.

The second aspect, mentioned in section (section 2.4), is the information that needs to be collected from any and all respondent households to a targeting survey. Social protection programmes with an economic wellbeing eligibility criterion verified via PMT are often targeted to specific population subgroups. For this purpose, the households are classified most commonly by demographic criteria, such as family structure or age group. The verification of such categorial eligibility implies the collection of a core set of relevant variables before income proxies are queried. Combined with items known up-front, such as geographic features, any data needed for administrative or statistical purposes makes up a core questionnaire administered to all respondents at the beginning of enumeration. We refer to the remainder of variables, which amount to 37 items in our Indonesian dataset, as optional. In the subsequent discussion and simulations, we take the core questions as given and refer to optional questions when discussing items such as questionnaire length.

Having been selected for administrative reasons, the short core questionnaire has limited predictive power for household consumption. At the other extreme, a full socio-economic population sample survey contains rich household information that confers greater predictive accuracy, but is unsuitable for administration to a large section of the population. The following three subsections describe methods that provide a menu of choices in between, and expose the relationship between survey cost and accuracy.

3.1 Truncated questionnaires

A simple approach is to order the optional features from most to least predictive. A truncated questionnaire, i.e. one that is limited to a certain number of questions, arises from adding the desired number of most predictive features to the core set. Computation of the predictive accuracy for each of the iteratively growing variable sets provides the cost-accuracy trade-off. In addition to a predictive model, this approach requires an ordering of features by their predictive power for household consumption. We test several variable selection methods to perform such an ordering and thus generate a sequence that can be truncated.

- *Stepwise selection.* The classic mechanism of PMT modelling[6]. To generate a complete variable sequence instead of a particular set of regressors, we adjust algorithm 1 by eliminating the stopping criterion and recording the order in which variables are added to the set of predictors.

- *Variable importance.* Tree-based predictive models perform variable selection during the construction of their tree structure. Our gradient boosting model, which is based on decision trees, generates two measures of variable importance. One is the proportion of trees in which each variable is used, the other is the sum of increases in the trees' objective function produced by the splits of each variable. We translate both the split and gain measures into two separate variable orderings. The first is to rank variables in descending order of either the number of splits, or of the gain they produce, when the full set of variables is considered jointly. The second uses a backward selection approach, recursively eliminating the variable with the lowest split or gain, respectively.

[6] Alternative methods have begun to gain traction among practitioners [Grosh et al., 2022].

- *Group lasso.* A variation of the Lasso, or Least Absolute Shrinkage and Selection Operator [Tibshirani, 1996], the group lasso's [Yuan and Lin, 2006] objective function allows regularization over grouped variables:

  $$min_{\beta}\frac{1}{2}||y-\sum_{l=1}^{m}X^{(l)}\beta^{(l)}||_2^2+\lambda\sum_{l=1}^{m}\sqrt{p_l}||\beta^{(l)}||_2$$

  where, following the exposition of Simon et al. [2013], $X^{(l)}$ is the submatrix of X with columns corresponding to the predictors in group l, $\beta^{(l)}$ is the coefficient vector of that group and p_l is the length of $\beta^{(l)}$. As in the basic lasso [Tibshirani, 1996], λ remains a free parameter that induces sparsity along the regularization path, only in this case jointly over groups and single variables. From the sets of groups selected for increasing values of λ, we construct a nested sequence of group sets that results in the variable sequence.

Note that we separate the variable selection approach that generates a question sequence from the predictive model that is used to predict along each sequence. The uniform prediction model used in this paper enables a comparison of the differences in length-accuracy trade-offs of the various sequences, as the starting accuracy of the core questionnaire and the final accuracy of the full variable set are identical across sequences. This clarifies the variable sequence that achieves the highest accuracy for a given number of questions (for the average household). Truncation is designed to achieve the best average result, but – unlike the following method – it fails to exploit the household-specific information that becomes available during enumeration.

3.2 Early stopping

In machine learning the term *early stopping* is commonly used to refer to a learning algorithm that stops training at a point where additional iterations yield no more benefit or lead to overfitting (e.g. Goodfellow et al. [2016]). In our setting, we re-purpose the term to refer to an interruption of household-level data collection when additional questions are expected to lack meaningful information about a household's consumption level. Like truncated questionnaires, this approach draws on variable sequence selection and gradient boosting as the predictive model. The key difference lies in the prediction of conditional quantiles instead of a conditional mean for each step along the variable sequence.

This change is accomplished by a change of objective function from mean squared error to quantile regression loss [Koenker and Bassett Jr, 1978]. By setting a symmetric prediction interval of a percentile θ and $(1-\theta)$, a pair of models estimates a plausible quantile range for a household's consumption level in view of the information accumulated at the current point of the question sequence. Enumeration proceeds until the prediction interval suggests that a household is either eligible or ineligible. Eligibility is assumed when the upper bound of the prediction interval lies below the programme's eligibility threshold, and ineligibility is assumed when the lower bound lies above. When early stopping is triggered in this way, the conditional mean[7] prediction for a model trained on the variables queried until this point is recorded as the household's final consumption estimate. Algorithm 2 summarizes the procedure, where X_{core}, X_{sel} and $X_{:,s}$ denote column-wise partitions of design matrix X, $f()$ is a prediction model for the conditional mean, q_θ is a quantile regression model for quantile θ.

Figure 3 illustrates the process for a single household. In this case, the prediction interval shifts down and becomes narrower as more information is collected. After 12 questions, the upper bound falls below the eligibility threshold, at which point the conditional mean of the household's consumption level is logged as the final estimate (dashed line), and enumeration ends. The trade-off between cost and accuracy is achieved via the width of the prediction interval determined by the percentile value θ. As θ is decreased, the prediction interval becomes narrower and is more likely to exclude the eligibility threshold. Early stopping will be triggered for a greater share of households, leading to a reduction in the average number of questions per household. A lower θ can be interpreted as a less certain prediction[8] that reduces enumeration cost at the expense of less precise consumption estimates.

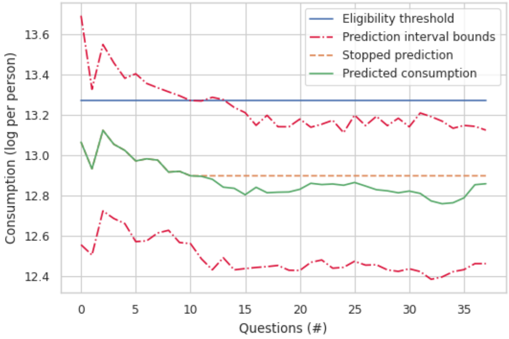


Figure 3: Illustration of early stopping for an example household

[7] The conditional median would be a point estimate alternative, and in a similar vein [Brown et al., 2018] showed that the use of quantile regression set to the intended coverage quantile of the target population can result in an exclusion error reduction.

[8] The predictive model does not calibrate the prediction intervals precisely to the true population quantile. As a result, a θ value of, say, 0.05 does not correspond accurately to a 95% probability of the true value being below the upper bound of the prediction interval, but it can be thought of as an approximation.

For truncated questionnaires, we noted that the ordering of variables by predictive power leads to diminishing accuracy gains as more items are added. Similarly, the prediction intervals that determine early stopping behaviour shift progressively less. This implies that households for which early stopping is not triggered early on are unlikely to breach the stopping threshold at any point. The result is a full questionnaire for those households close to the eligibility threshold, which raises average questionnaire length with little accuracy benefit. The following approach can overcome this issue.

Algorithm 2 Early stopping

```
Input: Training set $(\mathbf{X}, \mathbf{y})$, inference set $(\mathbf{X}^{nfr})$, core variables $\mathbf{X}_{core}$, variable sequence $\mathbf{s}$, eligibility threshold $\gamma$
  $\mathbf{X}_{sel} \leftarrow \mathbf{X}_{core}$                              ▷ Pre-compute models
  for each $s \in \mathbf{s}$ do:
      $\mathbf{X}_{sel} \leftarrow$ append$(\mathbf{X}_{:,s})$
      $f^s() \leftarrow$ fit$(\mathbf{X}_{sel}, \mathbf{y})$
      $q^s_\theta() \leftarrow$ fit$(\theta, \mathbf{X}_{sel}, \mathbf{y})$
      $q^s_{1-\theta}() \leftarrow$ fit$(1-\theta, \mathbf{X}_{sel}, \mathbf{y})$
  end for
  for each $\mathbf{x}_i \in \mathbf{X}^{nfr}$ do:                          ▷ Infer consumption
      $\mathbf{x}_{sel} \leftarrow \mathbf{X}^{nfr}_{i,core}$
      for each $s \in \mathbf{s}$ do:
          query$(\mathbf{X}_{i,s})$
          $\mathbf{x}_{sel} \leftarrow$ append$(\mathbf{x}_{sel}, \mathbf{X}_{i,s})$
          if $q^s_\theta(\mathbf{x}_{sel}) < \gamma$ or $q^s_{1-\theta}(\mathbf{x}_{sel}) > \gamma$ then:
              $y^s_i = f(\mathbf{x}_{sel})$
              break
          end if
      end for
      $y_i = f^s(\mathbf{x}_{sel})$
  end for
```

3.3 Truncated early stopping (TrESt)

TrESt performs early stopping, but with a limit on the maximum number of questions that households can be asked. This approach provides a household-specific stopping criterion for cases where eligibility status appears clear, while capping questionnaire length to balance cost and accuracy for more ambiguous cases. Figure 4 illustrates the concept, showing the EER-average length result for a range of truncations that emerge from the early stopping points for selected interval widths. Setting truncation to the full number of optional variables recovers the early stopping solution, whereas a maximum width prediction interval would yield the truncated questionnaire solution.

Truncated questionnaires and early stopping both expose the cost-accuracy trade-off via a single hyperparameter. For the former, this is the number of optional questions, and for the latter it is the width of prediction intervals. TrESt requires both hyperparameters, which can work in opposite directions; a longer questionnaire can be counteracted with a narrower prediction interval, and vice versa, which results in overlapping outcomes for different hyperparameter pairs. The TrESt solution consists of those hyperparameter pairs that generate the lower bound of accuracy-length outcomes (shown as a red line in fig. 8 in the results section). We use a separate validation set to identify the lower bound hyperparameters without overfitting, as cross-validation was already deployed on the training set in the generation of the variable sequence[9].

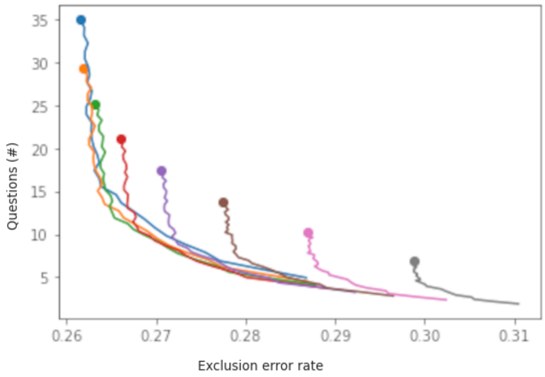


Figure 4: Illustrative truncation sequences for selected early stopping predictions. Colours represent different prediction intervals.

4 Results

This section presents the simulation results for each of the three proposed approaches, for a social protection programme with 40% population coverage in Indonesia. A PMT baseline provides a comparison with current policy practice. Whereas the PMT baseline consists of district-specific models, the other approaches are trained at national level with a district identifier in the core questionnaire that enables generation of locally adapted models. Our main outcome of interest is the number of questions that an approach requires versus the exclusion error rate. We favour the exclusion error rate (EER) as the evaluation metric most closely aligned with the targeting policy objective of identifying the eligible poor. Additional targeting metrics for a

[9] A more computationally intensive approach along the lines of Cawley and Talbot [2010] may be preferable in settings with limited data.

broader range of programme population coverages from 10% to 50% can be found in appendix A.

The graphical results show the number of optional questions that are required for a given accuracy, measured by EER. This ordering implies that policymakers choose a minimum required accuracy with reference to a benchmark method, rather than assigning a question budget per household. Recall that the full questionnaire starting point and the core questionnaire-only end point are shared, as the predictive model is identical across sequences. A method that is closer to the origin is preferable as it offers a superior cost-accuracy trade-off. The complete questionnaire that mirrors Indonesia's targeting survey of 37 optional questions results in a simulated EER of 26%[10]. At the other extreme, a gradient boosting model trained only on core questionnaire variables would achieve an EER of 37.78%. The range between core and full questionnaire spans almost half the EER's lower end value, implying a major – and likely unacceptable – decline in targeting accuracy when minimizing question numbers.

4.1 Truncated questionnaires

The first set of results is for questionnaires of uniform length administered to all households. We consider outcomes for sequences based on stepwise selection, the group lasso[11], and variable importance for the generation of grouped variables sequences that underlie these questionnaires. For the tree-based variable importance measures, we noted the option of using either split counts or loss function gains, as well as their computation via either a one-way ranking or a backwards selection-style recursive feature elimination. From the resulting four possible implementations, we show one-off gain as the best-performing one, which leaves recursive split-based elimination as a complement. Figure 5 shows the four methods' relative performance.

Figure 5 reveals stepwise selection to be the most effective approach in this setting, followed by the group lasso and then the variable importance measures. Whereas the performance penalty of the group lasso is moderate, the variable importance measures underperform significantly and cannot be recommended for this use case. Whereas there is a steep rise in the required number of questions to achieve the ultimate, small reductions in EER (top left), the sequencing of variables by predictive power results in a strong accuracy impact for initial questions (bottom right). The second key result is that fewer than 5 questions raise the EER strongly, and beyond 15 questions the accuracy gain becomes imperceptible for the best-performing method of stepwise selection. An intermediate range provides a moderate

[10] The full questionnaire would be selected when a gradient boosting model is used in the PMT algorithm instead of OLS on nation-wide data.

[11] We use the Group Lasso python library and the LightGBM framework [Ke et al., 2017] to implement gradient boosting.

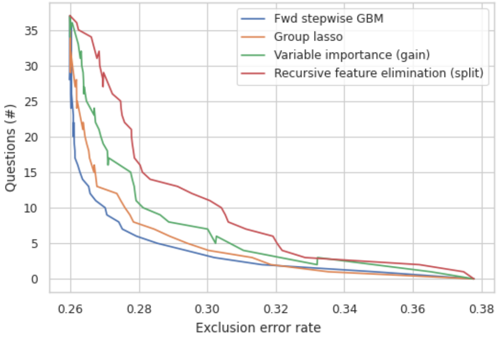


Figure 5: Truncated questionnaires: number of questions by EER comparison for several variables sequences

elasticity between EER and questionnaire length. It appears likely that policymakers with all but extreme preferences are likely to choose a point in this range for settings, where each question has a meaningful budget impact.

4.2 Early stopping

This subsection assesses whether early stopping, a basic adaptive method that processes household information during enumeration, can improve the cost-accuracy trade-off of a PMT-style approach. We compare outcomes among the same variable sequences as above, and then evaluate whether early stopping or truncated questionnaires yield better results in each case.

Recall that early stopping, which generates the length-accuracy trade-off via the width of prediction intervals, results in questionnaire length variation across households. The end points of the variable sequences are very similar, at 34.8 (variable importance/ gain), 34.5 (recursive feature elimination/ split), 34.5 (group lasso), and 34.2 (stepwise) questions, all resulting in an EER of 25.99%. The number of questions is below the full-length 37 as early stopping is triggered in some instances even when the interval width is set to a minimum of 0.01, as even basic information is sufficient to identify households with an extreme income level. The identical EER with that of a full questionnaire shows that there is no loss in accuracy from early stopping for that subset of households.

To constrain the range of outcomes to a statistically credible range, we have limited the maximum θ quantile to 0.3, resulting in prediction intervals limited by the 30th and the 70th percentiles at its most narrow. Note that the prediction bound collapses to the median as the quantile reaches 0.5 in our symmetric interval scheme. At that point, early stopping would be triggered at the core questionnaire for all households, yielding the same result as the truncation at zero questionnaire with EER 37.78% as above.

Accordingly, the end point of early stopping is the same across all approaches including truncated questionnaires, and the reader may picture the lines converging at this point.

The first main result, shown in fig. 6, mirrors that of the previous subsection: early stopping based on stepwise selection provides the best trade-off, followed the group lasso sequence and then the variable importance measures. The differences between variable selection approaches are notably smaller than for truncated questionnaires, suggesting that early stopping is less sensitive to the particular variable sequence used. The likely reason for this effect is that stopping happens at different points for different households, and also more often at more effective models, which smooths the outcomes.

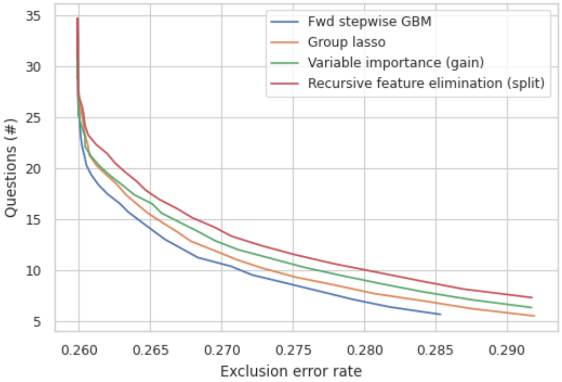


Figure 6: Early stopping: number of questions by EER comparison for several variable sequences

The shape of the stepwise selection curve shows that the trade-off between length and accuracy is remarkably similar. We observe that the elasticity becomes greater at around 20 questions. When plotted on the same graph, a direct comparison underlines the validity of this impression. Figure 7 compares the results of the truncation and early stopping approaches separately for each of the variable selection methods. Interestingly, the truncation vs early stopping results are nearly identical for both the stepwise selection and the group lasso plots, despite the substantial difference in the algorithms. The similarity suggests that both methods are able to exploit their respective variable sequences with similar efficacy.

In the case of variable importance, the early stopping approach provides better results than truncated questionnaires. We again interpret this relative outperformance as being related to a less precise ranking of variables by predictive power, and the smoothing across models that occurs in early stopping. Overall, the stepwise selection variable sequence retains the best length-accuracy result by a small margin over the group lasso. For this sequence, truncated questionnaires require slightly fewer questions for a given accuracy, when considering shorter questionnaires, whereas for long questionnaires the early stopping solution is more effective by what amounts to a negligible margin.

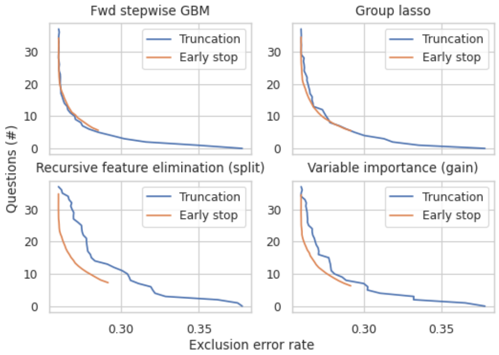


Figure 7: EER comparison of early stopping vs stepwise selection by number of questions

4.3 Truncated Early Stopping (TrESt)

Early stopping and truncated questionnaire can potentially be combined as they exploit the same variable sequence in distinct ways. The results shown in fig. 8 confirm the effectiveness of a joint approach in an empirical application[12]. The orange dots show the early stopping outcomes for various prediction interval widths, and the green line represents the range of truncated questionnaires. Grey lines emerging from each dot represent the truncation paths for the early stopping models, and their lower bound (in red) constitutes the TrESt solution.

The initial near-vertical decline of the grey lines confirms that the exclusion of the least predictive variables from questionnaires is similarly useful in an early stopping algorithm to reduce average questionnaire length. The zoomed-in section isolates the lower bound, and plots it against the current policy baseline to illustrate the performance difference. TrESt achieves a PMT-equivalent EER with an average of only 8.9 questions per household, compared with 22.4 for the PMT policy baseline. We review a key aspect of algorithm functioning before considering a full set of results across the different methods.

To appreciate the mechanics of the TrESt approach, consider the algorithm's behaviour for different pairs of questionnaire length and prediction interval width. Figure 9 shows the proportion of households for which early stopping is triggered along these dimensions. As expected, a larger quantile,

[12] Stepwise selection and a gradient boosting model are used to generate the TrESt algorithm results below.

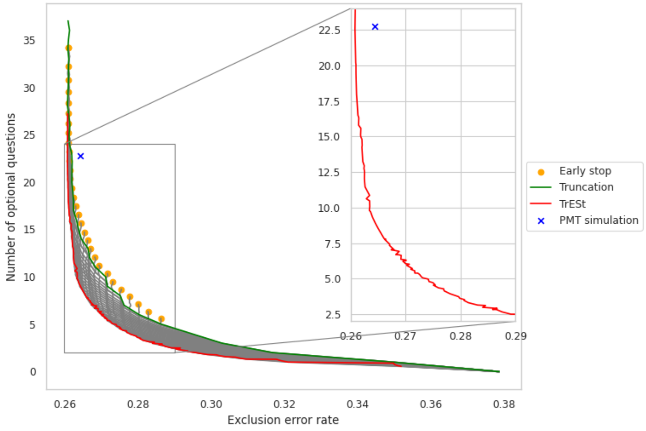


Figure 8: Overview plot

which results in a narrower prediction interval, increases the proportion of early stopping for a given questionnaire length. At the first percentile, there is practically no early stopping on completion of the core questionnaire, whereas at the 30th percentile this basic information is deemed sufficient to classify nearly half (45.3%) of households. At the other end, by the penultimate question, only 9.0% of households' prediction interval excludes the eligibility threshold for the 1st percentile, rising to 89.9% for the 30th percentile.

The information about a household's economic conditions in the first five optional questions has the strongest impact on stopping, as it shifts and narrows the prediction interval more than subsequent information. The share of households for which early stopping is triggered becomes increasingly flat thereafter. Even at the most sensitive end of the considered intervals (30th percentile), early stopping only increases by 1.5 percentage points over the last twenty questions. From truncated questionnaires, we already know that the consumption estimate's accuracy gain is very limited at this part of the questionnaire. Excluding the last variables from considerations thus has only a small impact on the ultimate consumption estimate, but it reduces questionnaire length by one-third.

4.4 Comparison table

Table 2 compares the results of the main approaches the main results in terms of the EER that corresponds to a particular questionnaire length. It demonstrates that a trade-off between average questionnaire length and targeting accuracy that can be generated with each of the three proposed methods. In our Indonesia simulation, both truncated questionnaires and

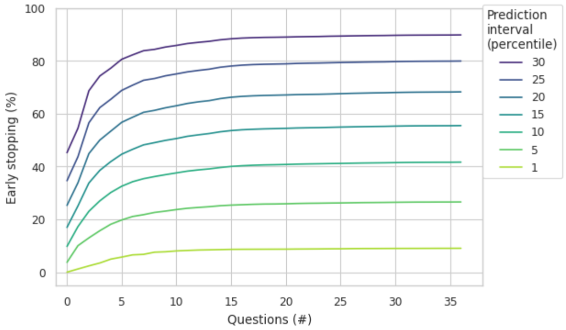


Figure 9: Proportion of stopped questionnaires

early stopping yielded remarkably similar outcomes. TrESt is able to leverage the respective mechanisms for a superior length-accuracy trade-off for the case of shorter questionnaires. Above circa 18 questions, early stopping becomes the most effective approach, albeit by a small margin[13]. For questionnaires with fewer than 20 items, at which point a notable trade-off between length and accuracy emerges, TrESt provides the best set of solutions. The resulting range of options begs the question which specification should be chosen, and we explore this issue for a specific setting in the following section.

Table 2: Comparison of EER% across models by questionnaire length (approximate length)

Approx. questions	PMT	Truncated	Early stop	TrESt
37	–	25.99	–	–
30	–	26.00	25.99 (29.5)	–
23	26.44 (22.8)	26.09	26.01 (23.2)	26.08 (23.3)
20	–	26.09	26.05 (20.3)	26.10 (19.8)
15	–	26.28	26.44 (14.7)	26.18 (15.1)
10	–	27.02	27.07 (10.3)	26.34 (9.9)
5	–	28.55	28.53 (5.6)	27.32 (5.0)
0	–	37.79	37.79	37.79

5 Case study: Urban Setting

The benefit of a method to allow much shorter questionnaires while maintaining accuracy becomes more tangible when placed into a particular policy

[13] The small underperformance of TrESt vs early stopping for long questionnaires can be attributed to the selection of hyperparameter pairs on the validation set.

context. This section calibrates a TrESt model to a particular urban setting, allowing us to weigh enumeration costs against estimation accuracy, and thereby maximize a policy objective. The following rudimentary policy simulation suggests scope for substantially improved targeting outcomes.

Three basic solutions for the cost-accuracy trade-off are apparent. One is maximum accuracy regardless of cost. Implicitly, the standard PMT falls into this category as it optimizes the model's predictive power regardless of questionnaire length. The other extreme would be the minimum cost solution of a core questionnaire. In many cases, the optimal solution may lie at a point where cost and accuracy are balanced to achieve a certain policy objective. Programme incidence, defined as the share of intended beneficiaries in the covered population, is a realistic objective, and in this case study we optimize for it on the assumption of a fixed budget. Accordingly, we select the TrESt model specification that yields the highest feasible incidence by weighing the extent of survey coverage against overall accuracy.

The setting we simulate is DKI Jakarta, Indonesia's capital city region that had about 10.5 million inhabitants within a densely populated urban environment. Information shared by the local government, which is charged with updating the social registry, suggests that enumerators who administer the current targeting questionnaire survey collect information from 10 to 20 households per day, and that each questionnaire takes between 15 and 30 minutes per household (personal communication, September 2021). Taking the midpoints of these ranges implies that the average travel time between households is 10 minutes. Based on the number of items that make up the core questionnaires and the implied travel time, the fixed and maximum variable times for each interview both amount to circa 16 minutes[14]. Assuming that each of the optional questions takes the same amount of time, variable time is reduced to the proportion of grouped questions out of the full questionnaire total.

The programme we simulate is the PBI-JKN subsidized health insurance programme intended to cover 40% of the population nationally, but which the DKI government has extended to 51% of households. Rounding to 50% to align with our simulations, and adjusting household figures accordingly, SUSENAS data indicate an incidence of 62.8% in 2019. A simplifying assumption at baseline is that all households which are included in the social registry receive the programme, as full coverage of all the surveyed households mirrors actual practice at national level, where the DTKS social registry contained the same 40% share of households that the programme aimed to cover [Pahlevi, 2019].

We simulate the impact of a shorter questionnaire by separating households into consumption deciles and splitting each decile into a surveyed and

[14] Based on 26 core items out of a total of 112 ungrouped variables, and rounding up to the full minute.

an unsurveyed group according to empirically observed percentages in the SUSENAS data. The PMT is distinct to the policy baseline as it would only require an average of 24.4 questions to be asked for the Jakarta area local models, whereas the status quo is based on the actual policy practice of querying a full questionnaire. A shorter questionnaire then allows additional households to be sampled at random from the unsurveyed population[15]. The additional households are surveyed with either a PMT or a TrESt questionnaire based on the stepwise selection sequence that proved most effective. To determine eligibility, households are ranked according to the PMT predictions, or the TrESt predictions that arise from the lower bound length-interval width pairs that yield its solution. Each hyperparameter pair results in an average number of questions, which in turn determines the total number of households that can be surveyed at a given accuracy. A corresponding incidence value is the main outcome for each pair.

Figure 10 shows the simulation results. The incidence curve is less smooth than for the national results due to a more limited number of observations for Jakarta. The chosen hyperparameter pair, marked by the red triangle, is a maximum questionnaire length of only 5 grouped items, alongside a medium level of early stopping sensitivity generated by θ set to 0.11. Note that this specification lies slightly below the best test set outcome as it was chosen according to its performance on a validation set.

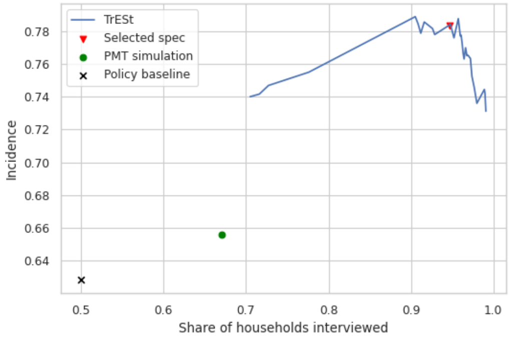


Figure 10: Simulated PBI-JKN incidence in DKI Jakarta at 50% coverage

The simulations indicate that a PMT that only queries the questions required for the model would allow for an additional 16.7% of the population to be surveyed. However, the linear model's modest predictive power in the DKI setting only raises the incidence to 65.6%. In contrast, TrESt achieves a significantly higher incidence at any model specification. At an average of 21.8 questions, even the model with the longest average questionnaire requires fewer questions on average than the PMT and still achieves

[15] Random selection is a conservative assumption, as households could be prioritized by likely eligibility, e.g. through poverty maps.

a considerably higher incidence of 74.1%. The chosen solution only uses an average of 3.9 questions per household, allowing 94.6% of the population to be sampled to achieve an incidence of 78.3%. Further average length reductions yield little benefit in terms of respondent numbers, and decrease accuracy so much that the overall effect on incidence is negative.

This case study suggests that TrESt has significant potential for raising incidence in a suitable setting. The case study relies on limited information and should be interpreted as a stylized example rather than as a fully calibrated use case. Nevertheless, it provides an illustration of the relationship between targeting survey coverage and household-level accuracy, and that it can be beneficial for desired policy outcomes to prioritize the former over the latter. Another caveat is that the low travel time between households in the urban setting under consideration here is a key environmental factor that promotes a close link between questionnaire length and incidence. Whether similar benefits would accrue where population density is lower could be assessed with enumeration meta-data for rural and peri-urban localities. Even when the number of questions is similar, the simulation suggests that the TrESt approach has a considerable advantage over a standard PMT, likely due to a more efficient orientation of survey resources to households with more uncertain eligibility status, and due to the greater predictive power of the non-linear, national-level machine learning model.

6 Discussion

We proposed a selection of practically feasible methods that expose a trade-off between accuracy and survey length, all based on variable sequences that order questions in order of predictive power. One simply truncates questionnaires to a certain length for all surveyed households, another deploys prediction intervals to generate a household-level early stopping criterion. For the most effective variable sequence – generated by stepwise selection – both approaches produce remarkably similar results in a simulation of a 40% programme coverage simulation with data from Indonesia, despite relying on different mechanisms. Compared with a policy baseline PMT that requires 22.8 questions to achieve an EER of 26.44%, truncation only needs 14 questions for a similar EER, and the early stopping method requires 18. A combined method that leverages their respective advantages achieves a superior length-accuracy trade-off at any point, and requires a mere 8.9 questions for an equivalent EER.

The methods presented here provide policymakers a choice on cost versus accuracy to suit their context and constraints. When selecting a point that maximizes incidence in a rudimentary simulation of an urban setting, a considerable improvement (to [78.3 %] incidence) appears feasible over both current practice ([65.6%]) and a more streamlined approach based on a PMT model ([62.8%]). Whether similar gains can be achieved in a less densely populated setting is an open question. Similarly, the assumptions of a uniform time per question and the ability to expand survey coverage to nearly the whole area population might prove overoptimistic in a real-world setting. Nevertheless, the urban case study highlights the potential benefit of expanding targeting survey coverage in some settings, even if it entails a moderate accuracy penalty.

A census sweep of the whole population with an extensive questionnaire would provide the most comprehensive data for accurate targeting. But if budgets are constrained, surveying more households with a shorter questionnaire may yield the best results. Beyond maximizing incidence with limited resources, policymakers would also enjoy greater flexibility through the methods presented here. The resources needed to conduct regular reassessment constrain targeting surveys to be conducted with multi-year gaps in the case of survey waves, and potentially also increase the re-assessment intervals in on-demand systems. A reduction in survey cost could contribute to a more dynamically supportive social protection system by raising the frequency of assessment and re-assessment. It is also worth noting that TrESt's efficient use of survey resources via household-level early stopping, a suitably truncated questionnaire, and a more effective predictive model appears to offer a clear accuracy improvement over the standard PMT for a given average questionnaire length.

To generate transparent results, our simulations consider a single programme with a fixed population coverage rate. Countries with social registries, of which Indonesia's DTKS is a well-documented example, usually target multiple programmes with differing population coverage rates that result in different eligibility thresholds. Prediction interval-based methods can be adjusted to this common scenario by using an eligibility interval instead of a threshold. The appropriate interval for each household can be identified according to characteristics queried in the core questionnaire. *Ceteris paribus*, wider eligibility intervals will result in longer questionnaires on average. A related caveat is that the targeting accuracy of new programmes that leverage existing consumption estimates but which have different eligibility thresholds – widely adopted during the recent COVID-19 pandemic – will be lower than that of estimates based on full questionnaires or newly collected data. Policymakers considering the introduction of a TrESt-style method are advised to simulate such contingencies in a fully calibrated model.

Implementation of the TrESt algorithm would require changes to enumeration practice. Software is one aspect, as it requires on-the-fly inference of prediction intervals for the stopping criterion. For Indonesia, an existing Android app used for the targeting survey could be adjusted for this purpose, but survey bodies elsewhere that currently rely on paper or off-the-shelf digital questionnaires would require new software. Deployment on sufficiently powerful hardware, such as a mid-range tablet or smartphone that can conduct the computations on the fly, or a reliable internet connection for processing in the cloud, would be required. Enumerator training is another implementation aspect that would require adjustment as the targeting survey questions need to be ordered by their stepwise regression sequence rather than the standard of thematic grouping. A final IT-related issue is that the social registry information collected with a TrESt or early stopping approach would collect jagged optional question data, so that survey designers would need to include all variables required for analytical or monitoring purposes in the core questionnaire.

Although the simulations presented here suggest scope for reducing survey costs while potentially reducing exclusion errors, this paper is only desk-based proof of concept. Indonesia has already achieved good targeting outcomes through years of investment in improving targeting processes and implementation. Continued improvements in design could further strengthen targeting outcomes at no additional cost, but piloting would be advisable to verify that the method is practical, that the results hold in the field, and also to collect additional data. For Indonesia, a simple PMT trial that collects the survey metadata to estimate realistic survey times would provide key budgetary inputs. Similarly, information on travel times between households would be important to calibrate time savings, particularly in rural districts and island locations where shorter questionnaires may only yield negligible savings. Detailed cost estimation along the lines of Fujii and van der Weide [2020] on the cost-effectiveness of double sampling would be advisable if implementation were to be considered.

One way to improve accuracy may be to draw on alternative data sources, such as the exploration of internet and phone expenditure for Indonesia's PMT by Pinxten [2021]. Such data can potentially support better classification, shorter questionnaires, or both. Similarly, alternative data preparation tailored to the early stopping approach may offer significant benefits in terms of accuracy or question numbers. On the other hand, the shortened questionnaire raises the risk of misreporting, as it concentrates both predictive power and enumerator attention on a few high-impact variables. A restriction to verifiable variables may mitigate this risk, but monitoring of response patterns would remain important to identify emerging subterfuge.

Beyond a more detailed assessment of financial and logistical aspects, additional applications with data from other countries would be helpful in assessing whether the early stopping algorithm can be a useful tool in other settings. The distributional impact in terms of unequal outcomes for different groups is another important aspect that would warrant further exploration. Appendix B outlines an illustrative group-level analysis which suggests that outcomes for pensioner households are similar for early stopping when compared with a standard PMT. Further reassurance would be gained by verification of estimation consistency for a full range of relevant vulnerable groups, as well as by a more extensive exploration of targeting fairness along the lines of Noriega-Campero et al. [2020]. The distributional analysis could be combined gainfully with a consideration of multidimensional poverty measures that assess impacts beyond the standard consumption-based perspective taken here.

In terms of methodological options, alternative variable selection methods that are less prone to suboptimal choices, different predictive models, or non-symmetric prediction intervals that align with the skew of consumption distributions may yield further improvements in the cost-accuracy trade-off. A promising research direction is to move beyond the stopping criterion to a survey design with a question sequence tailored on-the-fly to each household. The adaptation of the tree-based method in Bakker et al. [2021] from classification to regression would be one step in this direction. While computational constraints preclude deployment of a fully-adaptive approach in the field, the development of a semi-adaptive method that deploys a limited set of variable sequences may be a promising research avenue. A related direction would be to tailor data collection to the promotion of more equitable targeting outcomes for economically disadvantaged groups.

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Vivi Alatas, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. Targeting the poor: evidence from a field experiment in Indonesia. *American Economic Review*, 102(4):1206–40, 2012.

Ana Areias and Matthew Wai-Poi. Machine learning and prediction of beneficiary eligibility for social protection programs. In *Revisiting Targeting in Social Assistance: A New Look at Old Dilemmas*, chapter 8. World Bank, Washington DC, 2022.

Michiel A Bakker, Duy Patrick Tu, Humberto Riverón Valdés, Krishna P Gummadi, Kush R Varshney, Adrian Weller, and Alex Pentland. DADI: Dynamic Discovery of Fair Information with Adversarial Reinforcement Learning. *arXiv preprint arXiv:1910.13983*, 2019.

Michiel A Bakker, Duy Patrick Tu, Krishna P Gummadi, Alex Sandy Pentland, Kush R Varshney, and Adrian Weller. Beyond reasonable doubt: Improving fairness in budget-constrained decision making using confidence thresholds. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 346–356, 2021.

Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Sudarno Sumarto. The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia. *Journal of Public Economics Plus*, 1:100001, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning - limitations and opportunities*. self-published manuscript, 2021.

Caitlin Brown, Martin Ravallion, and Dominique Van de Walle. A Poor Means Test? Econometric targeting in Africa. *Journal of Development Economics*, 134:109–124, 2018.

Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Luc Christiaensen, Ethan Ligon, and Thomas Pave Sohnesen. Consumption subaggregates should not be used to measure poverty. *The World Bank Economic Review*, 2021.

David Coady, Margaret Grosh, and John Hoddinott. Targeting outcomes redux. *The World Bank Research Observer*, 19(1):61–85, 2004.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The Elements of Statistical Learning*. Number 10. Springer series in statistics, 2001.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Tomoki Fujii and Roy van der Weide. Is predicted data a viable alternative to real data? *The World Bank Economic Review*, 34(2):485–508, 2020.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Margaret Grosh and Judy L Baker. Proxy means tests for targeting social programs: simulations and speculation. *Living Standards Measurement Study Working Paper*, 118:1–49, 1995.

Margaret Grosh, Phillippe Leite, Matthew Wai-Poi, and Emil Tesliuc. *Revisiting Targeting in Social Assistance: A New Look at Old Dilemmas*. World Bank, Washington DC, 2022.

Robert M Groves and Steven G Heeringa. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3): 439–457, 2006.

Camilla Holmemo, Pablo Acosta, Tina George, Robert J Palacios, Juul Pinxten, Shonali Sen, and Sailesh Tiwari. Investing in People: Social Protection for Indonesia's 2045 Vision, 2020.

ILO. *World Social Protection Report 2020-22: Social protection at the crossroads – in pursuit of a better future.* International Labor Organization, 2021.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

S. Klasen, S. Lange, G. Hadiwijaja, J. Pinxten, and B.A. Wirapati. Evaluation of PMT-based targeting in Indonesia. Unpublished technical report. Technical report, World Bank, 2016.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Varun Kshirsagar, Jerzy Wieczorek, Sharada Ramanathan, and Rachel Wells. Household poverty classification in data-scarce environments: a machine learning approach. *arXiv preprint arXiv:1711.06813*, 2017.

Phillippe Leite, Tina George, Changqing Sun, Theresa Jones, and Kathy Lindert. Social registries for social assistance and beyond. 2017.

Kathy Lindert, Tina George Karippacheril, Inés Rodríguez Caillava, and Kenichi Nishikawa Chávez. *Sourcebook on the foundations of social protection delivery systems.* World Bank, Washington DC, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural information Processing Systems*, 30, 2017.

Linden McBride and Austin Nichols. Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550, 2018.

Paul Niehaus, Antonia Atanassova, Marianne Bertrand, and Sendhil Mullainathan. Targeting with agents. *American Economic Journal: Economic Policy*, 5(1):206–38, 2013.

Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A Bakker, Luis Tejerina, and Alex Pentland. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 241–251, 2020.

Tim Ohlenburg. Machine learning for proxy means testing in Indonesia. Unpublished technical report. Technical report, World Bank, 2020.

Franziska Ohnsorge and Shu Yu. The long shadow of informality. 2021.

Said Mirza Pahlevi. Indonesia's Unified Database (UDB). ADB Social Protection Week conference presentation, Nov 2019.

Utz Johann Pape and Johan A Mistiaen. Household expenditure and poverty measures in 60 minutes: a new approach with results from Mogadishu. *World Bank Policy Research Working Paper*, (8430), 2018.

Juul Pinxten. Estimating the improvement of predictive performance through inclusion of mobile phone and internet data expenditure variables in Indonesian PMT modelling. Unpublished technical report. 2021.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22 (2):231–245, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Achmad Tohari, Christopher Parsons, and Anu Rammohan. Targeting poverty under complementarities: Evidence from Indonesia's unified targeting system. *Journal of Development Economics*, 140:127–144, 2019.

World Bank. *Targeting Poor and Vulnerable Households in Indonesia.* World Bank, Washington DC, 2012.

World Bank. *Social Assistance Public Expenditure Review.* World Bank, Washington DC, 2017.

World Bank. Aspire: Atlas of social protection indicators of resilience and equity, 2022. URL `https://www.worldbank.org/en/data/datatopics/aspire`.

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

A Additional targeting metrics and coverage rates

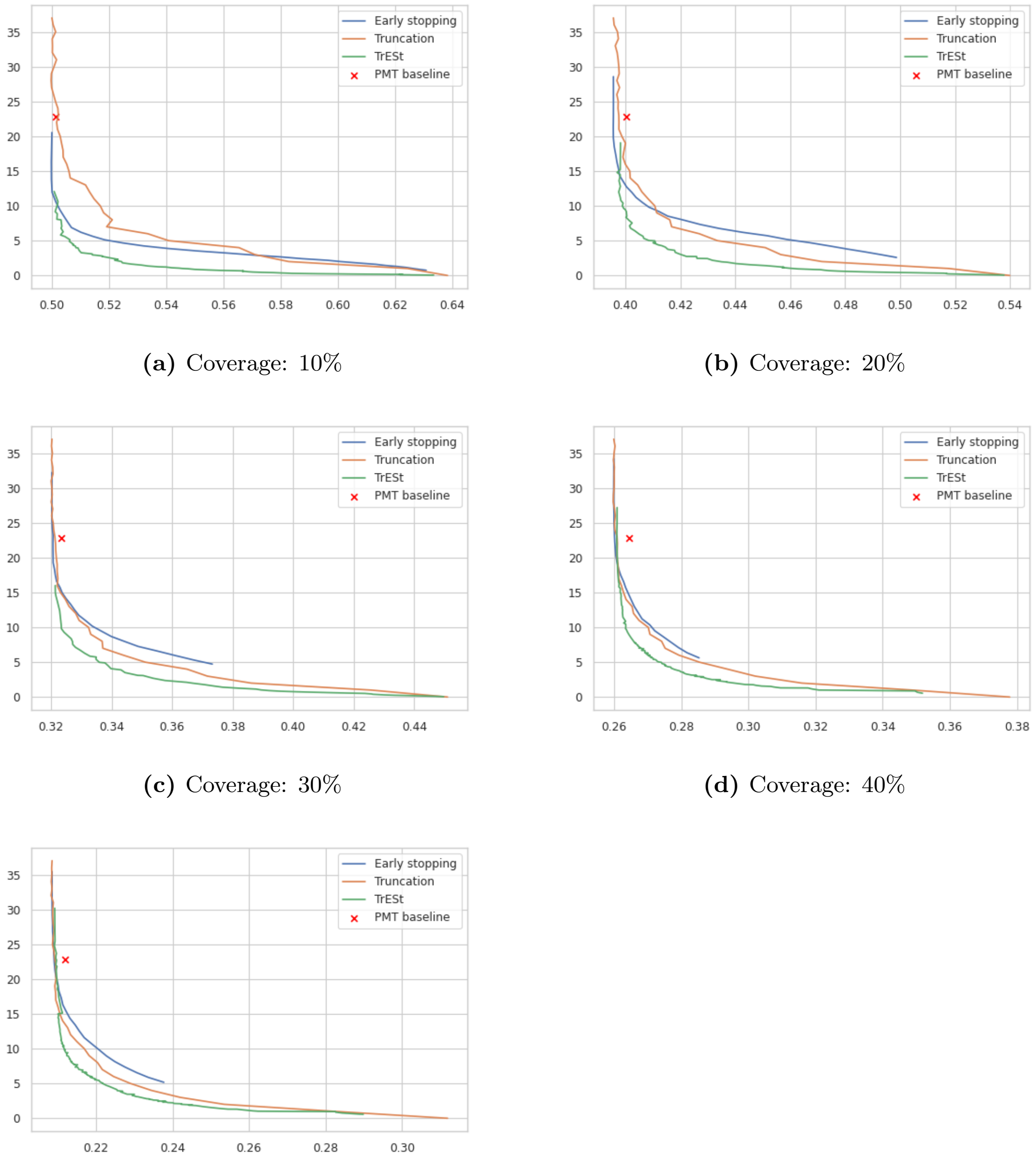
To show that the methods proposed perform well for a variety of programme coverage levels, this annex provides simulations for programmes that are targeted at 10%, 20%, 30%, 40%, and 50% of the population. Figure A1 shows the number of questions by EER across coverage levels. It highlights that the performance ranking of truncation and early stopping varies with the coverage level, which points to the need for context-specific selection if one of these is to be chosen. At low coverage rates, early stopping performs relatively better than truncation as the eligibility threshold moves to the lower bound and facilitates exclusion of non-eligible households with high consumption levels. TrESt not only maintains its superior performance, but outperforms truncation and early stopping by a wider margin than at the relatively elevated 40% coverage level considered in the main text.

Whereas the previous focus was on EER as the key targeting metric, this annex expands the results to a wider set of metrics. Table A1 outlines these, adding the commonly used inclusion error rate (IER), mean squared error (MSE), and the coefficient of variation (R^2). The results are shown in tables A2 to A5, for each of which we provide brief commentary below.

Table A1: Major targeting metrics

Metric	Description
Mean Squared Error (MSE)	The average squared difference between observed and predicted household consumption. MSE ranges over positive values and is denominated, and a smaller value signifies a more accurate predicted consumption level.
Coefficient of determination (R^2)	The proportion of variation in household consumption that is captured by the predictive model. A value of zero implies the lack of a linear relationship, whereas a value of one implies perfect correlation between predictor and predictand.
Exclusion Error Rate (EER)	The proportion of households with consumption below the eligibility threshold who are incorrectly predicted as being above. EER ranges between zero and one, and a smaller value signifies a more accurate predicted eligibility status.
Inclusion Error Rate (IER)	The proportion of households with consumption above the eligibility threshold who are incorrectly predicted as being below. IER also ranges between zero and one, and a smaller value signifies a more accurate predicted eligibility status.

The EER for each coverage level shown in table A2, in the same format as the main results table table 2, highlights that lower coverage rates result



(a) Coverage: 10%

(b) Coverage: 20%

(c) Coverage: 30%

(d) Coverage: 40%

(e) Coverage: 50%

Figure A1: Questionnaire length vs exclusion error rate by method for various programme coverage rates

in a higher proportion of exclusion among intended beneficiaries. At first sight, an EER of 50% may suggest that the eligibility assignment method is arbitrary, but random assignment would only capture 10% of beneficiaries and result in an EER of ca. 90%. Where household-level targeting is the policy preference for small coverage programmes, one can therefore argue that PMT, TrESt and other methods thus provide considerable advantages even if they only capture half of the intended beneficiaries. In terms of the number of questions needed to match PMT accuracy, around ten optional questionnaire items – and thus less than half the PMT outcome of around 23 – are needed for TrESt across coverage levels. For the Indonesia context, a questionnaire of this length would thus appear to provide a versatile method for most programmes.

Table A3 shows IER rates, a metric often considered important for political economy reasons. When the IER is defined with the number of beneficiaries in the denominator, e.g. as in Brown et al. [2018], then the fixed coverage rate used here would result in IER = EER. By using number of non-beneficiaries instead to align with local policy practice, we note that the inclusion error rate becomes IER = EER * coverage rate / (1 - coverage rate); EER and IER are then only equal in the case of 50% coverage. Given the close relationship between EER and IER in the fixed coverage regime considered here, the outcomes are qualitatively similar. As an effect of a growing denominator as the coverage level declines, the IER varies in the opposite direction to EER.

The coefficient of determination, or R^2, is a continuous targeting metric that is independent of the coverage level for methods such as the standard PMT and truncation. For early stopping and TrESt, the coverage level affects the algorithm's stopping criterion, and thus also feeds into estimation results. Table A4 shows that the point estimates of early stopping and TrESt are less precise than those of the PMT and truncation, which optimize for these point estimates. Although the classification-based EER and IER outcomes match the PMT at around ten questions, the R^2 at this point is considerably lower. However, this relative underperformance is of no concern if accurate eligibility is the ultimate objective. In fact, the lower point accuracy is a design feature as the algorithms stop trying to pinpoint the consumption level as soon as they identify a sufficient estimated difference to the eligibility threshold. As such, inferior continuous targeting metrics are unproblematic as long as there is no separate need for accurate consumption estimates.

The same insights hold for the MSE results shown in table A5 as for the previous table (not least because, as IER and EER, they are mathematically related). The PMT and truncation display the same MSE across coverage levels, and truncation achieves a similar result with ca. 10 versus 22.8 questions. Early stopping and TrESt drop off due to their mechanics, and can only match the PMT in question numbers for its MSE level, but not trun-

Table A2: Exclusion error % by model and number of question for various coverage rates (approx. no. of questions in brackets)

Number of questions	PMT	Truncated	Early stop	TrESt
10% coverage				
37.0	–	50.01 (37.0)	–	–
30.0	–	50.08 (30.0)	–	–
22.8	50.16 (22.8)	50.23 (23.0)	–	–
20.0	–	50.30 (20.0)	50.01 (20.5)	–
15.0	–	50.60 (15.0)	–	50.09 (12.1)
10.0	–	51.70 (10.0)	50.15 (10.5)	50.21 (10.3)
5.0	–	54.09 (5.0)	51.85 (5.1)	50.62 (4.9)
0.0	–	63.82 (0.0)	63.08 (0.8)	63.34 (0.1)
20% coverage				
37.0	–	39.56 (37.0)	–	–
30.0	–	39.76 (30.0)	–	–
22.8	40.05 (22.8)	39.76 (23.0)	39.56 (23.0)	–
20.0	–	39.87 (20.0)	39.56 (19.8)	39.82 (19.1)
15.0	–	40.16 (15.0)	39.75 (15.3)	39.70 (14.8)
10.0	–	41.07 (10.0)	40.86 (9.8)	39.89 (10.0)
5.0	–	43.35 (5.0)	45.88 (5.2)	40.86 (5.0)
0.0	–	53.99 (0.0)	–	53.79 (0.1)
30% coverage				
37.0	–	32.03 (37.0)	–	–
30.0	–	32.03 (30.0)	32.03 (29.7)	–
22.8	32.34 (22.8)	32.12 (23.0)	32.06 (23.2)	–
20.0	–	32.17 (20.0)	32.07 (20.2)	–
15.0	–	32.31 (15.0)	32.37 (14.9)	32.14 (15.0)
10.0	–	33.24 (10.0)	33.36 (10.2)	32.34 (9.8)
5.0	–	35.11 (5.0)	37.33 (4.7)	33.59 (5.0)
0.0	–	45.10 (0.0)	–	44.97 (0.1)
40% coverage				
37.0	–	25.99 (37.0)	–	–
30.0	–	26.00 (30.0)	25.99 (29.5)	26.09 (27.2)
22.8	26.44 (22.8)	26.09 (23.0)	26.01 (23.2)	26.08 (22.6)
20.0	–	26.09 (20.0)	26.05 (20.3)	26.10 (20.1)
15.0	–	26.28 (15.0)	26.44 (14.7)	26.19 (15.0)
10.0	–	27.02 (10.0)	27.07 (10.3)	26.35 (10.1)
5.0	–	28.55 (5.0)	28.53 (5.6)	27.41 (5.0)
0.0	–	37.79 (0.0)	–	35.19 (0.5)
50% coverage				
37.0	–	20.85 (37.0)	–	–
30.0	–	20.86 (30.0)	20.85 (30.5)	20.91 (30.2)
22.8	21.17 (22.8)	20.91 (23.0)	20.90 (22.5)	20.96 (22.7)
20.0	–	20.95 (20.0)	20.95 (20.3)	20.98 (19.8)
15.0	–	21.05 (15.0)	21.21 (15.4)	21.00 (15.0)
10.0	–	21.69 (10.0)	22.09 (9.8)	21.17 (9.9)
5.0	–	22.89 (5.0)	23.77 (5.2)	22.16 (5.0)
0.0	–	31.19 (0.0)	–	28.99 (0.6)

Table A3: Inclusion error % by model and number of question for various coverage rates (approx. no. of questions in brackets)

	PMT	Truncated	Early stop	TrESt
10% coverage				
37.0	–	5.56 (37.0)	–	–
30.0	–	5.56 (30.0)	–	–
22.8	5.57 (22.8)	5.58 (23.0)	–	–
20.0	–	5.59 (20.0)	5.56 (20.5)	–
15.0	–	5.62 (15.0)	–	5.57 (12.1)
10.0	–	5.74 (10.0)	5.57 (10.5)	5.58 (10.3)
5.0	–	6.01 (5.0)	5.76 (5.1)	5.62 (4.9)
0.0	–	7.09 (0.0)	7.01 (0.8)	7.04 (0.1)
20% coverage				
37.0	–	9.89 (37.0)	–	–
30.0	–	9.94 (30.0)	–	–
22.8	10.01 (22.8)	9.94 (23.0)	9.89 (23.0)	–
20.0	–	9.97 (20.0)	9.89 (19.8)	9.96 (19.1)
15.0	–	10.04 (15.0)	9.94 (15.3)	9.92 (14.8)
10.0	–	10.27 (10.0)	10.21 (9.8)	9.97 (10.0)
5.0	–	10.84 (5.0)	11.47 (5.2)	10.21 (5.0)
0.0	–	13.50 (0.0)	–	13.45 (0.1)
30% coverage				
37.0	–	13.73 (37.0)	–	–
30.0	–	13.73 (30.0)	13.73 (29.7)	–
22.8	13.86 (22.8)	13.77 (23.0)	13.74 (23.2)	–
20.0	–	13.79 (20.0)	13.74 (20.2)	–
15.0	–	13.85 (15.0)	13.87 (14.9)	13.77 (15.0)
10.0	–	14.24 (10.0)	14.30 (10.2)	13.86 (9.8)
5.0	–	15.05 (5.0)	16.00 (4.7)	14.40 (5.0)
0.0	–	19.33 (0.0)	–	19.27 (0.1)
40% coverage				
37.0	–	17.33 (37.0)	–	–
30.0	–	17.33 (30.0)	17.33 (29.5)	17.39 (27.2)
22.8	17.63 (22.8)	17.39 (23.0)	17.34 (23.2)	17.39 (22.6)
20.0	–	17.39 (20.0)	17.37 (20.3)	17.40 (20.1)
15.0	–	17.52 (15.0)	17.63 (14.7)	17.46 (15.0)
10.0	–	18.01 (10.0)	18.04 (10.3)	17.57 (10.1)
5.0	–	19.03 (5.0)	19.02 (5.6)	18.28 (5.0)
0.0	–	25.19 (0.0)	–	23.46 (0.5)
50% coverage				
37.0	–	20.85 (37.0)	–	–
30.0	–	20.86 (30.0)	20.85 (30.5)	20.91 (30.2)
22.8	21.17 (22.8)	20.91 (23.0)	20.90 (22.5)	20.96 (22.7)
20.0	–	20.95 (20.0)	20.95 (20.3)	20.98 (19.8)
15.0	–	21.05 (15.0)	21.21 (15.4)	21.00 (15.0)
10.0	–	21.69 (10.0)	22.09 (9.8)	21.17 (9.9)
5.0	–	22.89 (5.0)	23.77 (5.2)	22.16 (5.0)
0.0	–	31.19 (0.0)	–	28.99 (0.6)

Table A4: Coefficient of determination (R^2) by model and number of question for various coverage rates (approx. no. of questions in brackets)

	PMT	Truncated	Early stop	TrESt
10% coverage				
37.0	–	0.63 (37.0)	–	–
30.0	–	0.63 (30.0)	–	–
22.8	0.62 (22.8)	0.63 (23.0)	–	–
20.0	–	0.63 (20.0)	0.55 (20.5)	–
15.0	–	0.63 (15.0)	–	0.46 (12.1)
10.0	–	0.61 (10.0)	0.44 (10.5)	0.46 (10.3)
5.0	–	0.57 (5.0)	0.37 (5.1)	0.43 (4.9)
0.0	–	0.31 (0.0)	0.32 (0.8)	0.31 (0.1)
20% coverage				
37.0	–	0.63 (37.0)	–	–
30.0	–	0.63 (30.0)	–	–
22.8	0.62 (22.8)	0.63 (23.0)	0.56 (23.0)	–
20.0	–	0.63 (20.0)	0.53 (19.8)	0.55 (19.1)
15.0	–	0.63 (15.0)	0.49 (15.3)	0.51 (14.8)
10.0	–	0.61 (10.0)	0.43 (9.8)	0.51 (10.0)
5.0	–	0.57 (5.0)	0.38 (5.2)	0.46 (5.0)
0.0	–	0.31 (0.0)	–	0.31 (0.1)
30% coverage				
37.0	–	0.63 (37.0)	–	–
30.0	–	0.63 (30.0)	0.61 (29.7)	–
22.8	0.62 (22.8)	0.63 (23.0)	0.56 (23.2)	–
20.0	–	0.63 (20.0)	0.54 (20.2)	–
15.0	–	0.63 (15.0)	0.49 (14.9)	0.55 (15.0)
10.0	–	0.61 (10.0)	0.45 (10.2)	0.55 (9.8)
5.0	–	0.57 (5.0)	0.40 (4.7)	0.50 (5.0)
0.0	–	0.31 (0.0)	–	0.31 (0.1)
40% coverage				
37.0	–	0.63 (37.0)	–	–
30.0	–	0.63 (30.0)	0.61 (29.5)	0.61 (27.2)
22.8	0.62 (22.8)	0.63 (23.0)	0.57 (23.2)	0.60 (22.6)
20.0	–	0.63 (20.0)	0.55 (20.3)	0.58 (20.1)
15.0	–	0.63 (15.0)	0.51 (14.7)	0.56 (15.0)
10.0	–	0.61 (10.0)	0.48 (10.3)	0.55 (10.1)
5.0	–	0.57 (5.0)	0.44 (5.6)	0.51 (5.0)
0.0	–	0.31 (0.0)	–	0.38 (0.5)
50% coverage				
37.0	–	0.63 (37.0)	–	–
30.0	–	0.63 (30.0)	0.61 (30.5)	0.62 (30.2)
22.8	0.62 (22.8)	0.63 (23.0)	0.58 (22.5)	0.60 (22.7)
20.0	–	0.63 (20.0)	0.57 (20.3)	0.60 (19.8)
15.0	–	0.63 (15.0)	0.54 (15.4)	0.60 (15.0)
10.0	–	0.61 (10.0)	0.49 (9.8)	0.56 (9.9)
5.0	–	0.57 (5.0)	0.45 (5.2)	0.51 (5.0)
0.0	–	0.31 (0.0)	–	0.39 (0.6)

cation. As such, a truncation approach may be preferable in settings where survey costs should be minimized, but accurate consumption estimates are required for uses outside the programme's eligibility determination mechanism.

Table A5: Mean squared error by model and number of question for various coverage rates (approx. no. of questions in brackets)

	PMT	Truncated	Early stop	TrESt
10% coverage				
37.0	–	0.16 (37.0)	–	–
30.0	–	0.16 (30.0)	–	–
22.8	0.17 (22.8)	0.16 (23.0)	–	–
20.0	–	0.16 (20.0)	0.19 (20.5)	–
15.0	–	0.16 (15.0)	–	0.23 (12.1)
10.0	–	0.17 (10.0)	0.24 (10.5)	0.23 (10.3)
5.0	–	0.18 (5.0)	0.27 (5.1)	0.25 (4.9)
0.0	–	0.30 (0.0)	0.30 (0.8)	0.30 (0.1)
20% coverage				
37.0	–	0.16 (37.0)	–	–
30.0	–	0.16 (30.0)	–	–
22.8	0.17 (22.8)	0.16 (23.0)	0.19 (23.0)	–
20.0	–	0.16 (20.0)	0.20 (19.8)	0.19 (19.1)
15.0	–	0.16 (15.0)	0.22 (15.3)	0.21 (14.8)
10.0	–	0.17 (10.0)	0.25 (9.8)	0.21 (10.0)
5.0	–	0.18 (5.0)	0.27 (5.2)	0.24 (5.0)
0.0	–	0.30 (0.0)	–	0.30 (0.1)
30% coverage				
37.0	–	0.16 (37.0)	–	–
30.0	–	0.16 (30.0)	0.17 (29.7)	–
22.8	0.17 (22.8)	0.16 (23.0)	0.19 (23.2)	–
20.0	–	0.16 (20.0)	0.20 (20.2)	–
15.0	–	0.16 (15.0)	0.22 (14.9)	0.19 (15.0)
10.0	–	0.17 (10.0)	0.24 (10.2)	0.20 (9.8)
5.0	–	0.18 (5.0)	0.26 (4.7)	0.22 (5.0)
0.0	–	0.30 (0.0)	–	0.30 (0.1)
40% coverage				
37.0	–	0.16 (37.0)	–	–
30.0	–	0.16 (30.0)	0.17 (29.5)	0.17 (27.2)
22.8	0.17 (22.8)	0.16 (23.0)	0.19 (23.2)	0.17 (22.6)
20.0	–	0.16 (20.0)	0.20 (20.3)	0.18 (20.1)
15.0	–	0.16 (15.0)	0.21 (14.7)	0.19 (15.0)
10.0	–	0.17 (10.0)	0.23 (10.3)	0.19 (10.1)
5.0	–	0.18 (5.0)	0.24 (5.6)	0.21 (5.0)
0.0	–	0.30 (0.0)	–	0.27 (0.5)
50% coverage				
37.0	–	0.16 (37.0)	–	–
30.0	–	0.16 (30.0)	0.17 (30.5)	0.17 (30.2)
22.8	0.17 (22.8)	0.16 (23.0)	0.18 (22.5)	0.17 (22.7)
20.0	–	0.16 (20.0)	0.19 (20.3)	0.18 (19.8)
15.0	–	0.16 (15.0)	0.20 (15.4)	0.17 (15.0)
10.0	–	0.17 (10.0)	0.22 (9.8)	0.19 (9.9)
5.0	–	0.18 (5.0)	0.24 (5.2)	0.21 (5.0)
0.0	–	0.30 (0.0)	–	0.27 (0.6)

B Example analysis of group-level outcome differences

Changes in policy implementation, such as the use of a new algorithm, may have distributional effects in the sense of systematic outcome differences across beneficiary groups. Most commonly, groups are defined by demographic or ethnic characteristics, but any economically vulnerable minority would be a useful unit of analysis. The machine learning literature refers to group-based outcome issues as fairness (see Barocas et al. [2021] for a textbook-style overview), with various definitions of what constitutes a fair outcome at group level. While a comprehensive assessment of the fairness effects of PMT vs early stopping is beyond the scope of this paper, this annex provides a simple example of the kind of analysis that could be conducted to ensure equitable policy outcomes across the population.

We select two-person pensioner households as an example of an economically vulnerable group. The histogram in Figure A2 shows the EER distribution for this group across districts, both for the PMT (red) and the early stopping (blue) algorithm. The purple area is the overlap between the group's outcomes for the two estimators, while the dotted lines show the mean respective EER rates. The means lie close together at 25.6% for the PMT and 25.8% for early stopping, and the predominance of purple area also implies that there is only a marginal increase in misclassification risk for this group when the early stopping algorithm is used instead of the PMT baseline. There is slightly more dispersion for early stopping, evident in the somewhat higher frequency of districts in which either all or none of the eligible pensioner households are classified correctly. Apart from this, the overall pattern is as similar as can be expected for a statistical process subject to randomness; we can conclude that the early stopping algorithm does not produce disparate outcomes for this group.

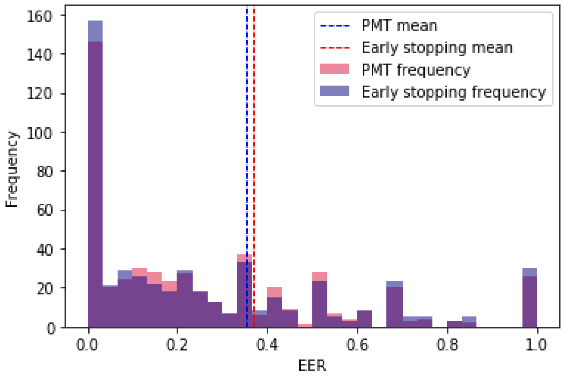


Figure B1: EER in a 40% coverage programme for two-person pension age households, early stopping vs PMT models