Policy Research Working Paper 8142

Predicting School Dropout with Administrative Data

New Evidence from Guatemala and Honduras

Melissa Adelman
Francisco Haimovich
Andres Ham
Emmanuel Vazquez

WORLD BANK GROUP
Education Global Practice Group
July 2017

Policy Research Working Paper 8142

Abstract

Across Latin America, school dropout is a growing concern, because of its negative social and economic consequences. Although a wide range of interventions hold potential to reduce dropout rates, policy makers in many countries must first address the basic question of how to target limited resources effectively for such interventions. Identifying who is most likely to drop out and, therefore, who should be prioritized for targeting, is a prediction problem that has been addressed in a rich set of research in countries with strong education system data. This paper makes use of newly established administrative data systems in Guatemala and Honduras, to estimate some of the first dropout prediction models for lower-middle-income countries. These models can correctly identify 80 percent of sixth grade students who will drop out in the transition to lower secondary school, performing as well as models used in the United States and providing more accurate results than other commonly used targeting approaches.

This paper is a product of the Education Global Practice Group. 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://econ.worldbank.org. The authors may be contacted at madelman@worldbank.org and fhaimovich@worldbank.org.

Produced by the Research Support Team

Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras

Melissa Adelman, Francisco Haimovich, Andres Ham, and Emmanuel Vazquez*

Keywords: Dropout, Prediction, Early Warning, Guatemala, Honduras
JEL Classification: I20, I21, C51, C53

*The authors thank the Ministry of Education of Guatemala (*Ministerio de Educación* – MINEDUC) and the Secretary of Education of Honduras (*Secretaría de Educación* – SEDUC) for providing access to the anonymized administrative data on which this study is based.

1. Introduction

As a result of sustained public and private investment in recent decades, primary net enrollment rates across all low and middle-income countries now average 91%, while secondary and tertiary enrollment rates are growing quickly. In the Latin America and Caribbean region, primary enrollment is approaching universality, and secondary net enrollment has grown 50% in the last two decades to over 75% (United Nations 2015; World Bank 2016). For many countries, however, attainment continues to fall short of aspirations, as high rates of enrollment in early grades quickly decline due to students dropping out before completing a full course of basic education (Bassi, Busso, and Muñoz 2016). For example, in Guatemala and Honduras, the countries of focus in this paper, education is *de jure* compulsory through ninth grade, but nearly 40% of sixth graders drop out before getting there. For young people who drop out prematurely, global evidence suggests that, on average, they will earn less and experience more social and economic challenges than their peers with more years of completed education (Patrinos and Psacharopoulos 2004; Oreopoulos and Salvanes 2011; Bentaouet-Kattan and Székely 2014).

Across Latin America, school dropout is a pipeline for expanding the population of underskilled and underengaged youth, contributing to social and economic challenges (Cardenas, De Hoyos, and Székely 2015). Several branches of research from North America, Latin America, and other regions have focused on identifying the causes of dropout, and point to multiple, interacting factors that affect learning, progression through grades, and ultimately dropout. Economics literature on dropout starts from the foundational human capital theory, in which individuals decide whether to persist or drop out of school by weighing the marginal expected costs of continuing investment in education (such as school fees and the opportunity cost of lost earnings) against the marginal expected benefits of acquiring more years of schooling (such as increased future wages from higher skills) (Becker 1967). Several authors have identified a range of individual, community, and broader factors that help shape these costs and benefits for each student, as well as the decision-making processes used to compare them, such as household wealth, schooling quality, and labor market conditions (Behrman, de Hoyos, and Szekely 2015; Adelman and Székely 2016). A large education literature focuses on dropout as the ultimate outcome of a process of disengagement from school, and demonstrates that dropouts can be grouped into distinct typologies, based on the factors driving their decision (Fortin et al 2006; Ananga 2011; Bowers and Sprott 2012).

Consequently, a range of interventions can be effective at reducing dropout and increasing attainment depending on the context, from conditional cash transfers to providing information on returns to schooling, to training on socio-emotional skills (Barrera-Osorio et al 2011; Nguyen 2008; Jensen 2010; Fryer 2013; Avitabile and de Hoyos 2015; Heller et al 2016).

While understanding why students drop out is critical, the ability of policy makers to respond effectively depends on answering an even more fundamental question – who is most likely to drop out? This question may appear relatively easy to answer, particularly in countries, regions, or localities with high dropout rates, as one might assume dropouts are concentrated in particularly disadvantaged or dysfunctional schools, or among students with particular characteristics. However, dropouts are often spread throughout schools and not readily identifiable by single characteristics, reflecting the complexity of the issue as documented in the dropout typology literature mentioned above. For example, in Guatemala, one of the countries in this study, over half of sixth grade students who drop out in the transition to lower secondary are spread across 70% of the country's primary schools, where the dropout rate is below 50%, and while 50% of students who score in the lowest quartile of a sixth-grade standardized exam drop out, so do 20% of those who score in the highest quartile.

Accurately identifying students at risk of dropping out in order to target effective interventions to where they are most needed is particularly important in contexts of limited resources and competing priorities, which describes most of the world's education systems. Largely based on the rich administrative data available in many U.S. school systems, research on dropout prediction is providing an increasingly sound empirical base for accurately predicting who will drop out several months to several years before dropout occurs, and over 30 U.S. states currently have in place some form of "early warning" system (O'Cummings and Therriault 2015). Similarly, the majority of European countries report monitoring "early warning" signs of potential dropout through their management information systems, primarily at the school level, but in some cases nationally (European Commission 2013). For many middle-income countries, which have invested in setting up information management systems in recent years, answering the prediction problem of dropout is now becoming possible through the use of consistent student-level data. In both Guatemala and Honduras, for example, student and school-level data are now digitalized in networked administrative databases, including unique student identifiers that allow tracking students over time, and, in the case of Guatemala, that can be directly linked to standardized test data.

In this paper, we make use of these administrative data to estimate early warning models of dropout in primary and secondary school.[1] Using linear regressions and basic prediction concepts, we are able to accurately predict approximately 80% of the sixth-grade students who dropped out within the next year in Guatemala and Honduras, performing at comparable levels to models used in the U.S. These early warning models, which are based on routinely collected data and relatively simple analytical techniques, are feasible to implement in a wide range of country contexts. By providing an accurate means of targeting, these models could substantially reduce the misallocation of program resources: in a simple simulation of a modest dropout prevention program, targeting students based on these models rather than targeting poor municipalities or high-dropout schools could reduce misallocation of resources by 30 to 80%.

The remainder of the paper is organized as follows. Section 2 briefly reviews the concepts and empirical evidence on early warning systems, which primarily come from the U.S. Section 3 describes the data and the current dropout situation in both Guatemala and Honduras, while Section 4 presents the results of constructing early warning models. Section 5 concludes with a brief summary of the findings and suggestions for future research.

2. Predicting dropout: Concepts and evidence

The quantitative literature on dropout prediction stretches back at least 30 years in the U.S. and Canada, and includes a variety of methods ranging from using single variables (often called “dropout flags”) to applied statistical learning (or “machine learning”) based on large data sets (in terms of both number of variables and number of observations). This research has been primarily concerned with using readily available data to provide accurate predictions to school system managers on who is most likely to drop out prior to completing a full course of primary or secondary education. The variables used for prediction are based on both the conceptual

[1] The authors thank the Ministry of Education of Guatemala (*Ministerio de Educación* –MINEDUC) and the Secretary of Education of Honduras (*Secretaría de Educación* – SEDUC) for providing access to the anonymized administrative data on which this study is based.

underpinnings discussed in the previous section and on the types of data usually available in school systems' administrative information systems. This means that in practice most dropout prediction is based on indicators of students' engagement in school and learning – such as attendance, behavioral infractions, course grades, and exam performance – as well as some socioeconomic characteristics – such as race and qualification for support programs (e.g. free lunch).

While most papers on the subject assert that their preferred method yields accurate predictions, the literature has been limited by use of inconsistent language and metrics for defining accuracy. In a 2013 systematic review, Bowers, Sprott, and Taff lay out a common set of prediction concepts and measure the existing literature against these metrics, finding substantial variation in the accuracy of prediction across methods. Below, we borrow from their review to describe the prediction concepts upon which our results are based.

The accuracy of predicting a binary outcome like dropout is illustrated by the simple event table in Table 1 (Bowers, Sprott, and Taff 2013; Stuit et al 2016). In order to provide useful information to policy makers attempting to target resources, the objective of most dropout prediction is to maximize sensitivity (the share of eventual dropouts correctly identified), while minimizing the rate of false positives (students identified as likely to drop out who in fact will not) (Gleason and Dynarski 2002). The rates of false positives and false negatives (students identified as likely to graduate who in fact will drop out) corresponds to the concepts of Type I and Type II errors, respectively, in statistical hypothesis testing (Sheskin 2004; Rogosa 2005).

These two objectives, of maximizing sensitivity and minimizing false positives, may be at odds with each other, as predictions that identify the majority of dropouts can often also incorrectly classify many students as dropouts. For example, several studies included in Bowers, Sprott, and Taff's review of single dropout flags achieve sensitivities of over 80% (e.g., correctly predicting over 80% of eventual dropouts), but have false positive rates over 50% (e.g., incorrectly predicting that over 50% of eventual graduates will drop out). The tradeoff between these two objectives can be measured using a Receiver Operating Characteristic (ROC) curve from the signal detection literature, which simply plots sensitivity (or true-positive proportion) against the false-positive proportion in x-y space for a given dropout prediction measure (Bowers, Sprott, and Taff 2013).

As Figure 1 illustrates, perfect dropout prediction would correctly classify all eventual dropouts and graduates, while random guessing would on average produce an equal number of false positives and false negatives. For any given dropout prediction measure, a ROC curve can be plotted based on the cutoff value set to define who is classified as a future dropout, and the optimal cutoff value determined through one of several related methods to trade off sensitivity and false-positives, including by setting the value at the point closest to perfect prediction (Fluss, Faraggi, and Reiser 2005; Liu 2012).

Based on these concepts of prediction accuracy, Bowers, Sprott, and Taff (2013) find that a large share of dropout predictors (across methodologies) perform quite badly, either identifying only a very small share of all eventual dropouts (clustering in the bottom right-hand corner of the ROC graph) or being only marginally better than random guessing (close to the 45-degree line). The most accurate predictors in their review all use growth mixture modeling, a form of multivariate longitudinal analysis that follows the academic trajectory of individual students, for example modeling learning outcomes over time (Muthen 2004; Janosz et al 2008; Bowers and Sprott 2012). These prediction methods achieve 80-90% sensitivity (correctly identifying 80-90% of eventual dropouts as being likely to drop out), and 10-20% false-positive rates (incorrectly identifying 10-20% of eventual graduates as being likely to drop out). The relatively strong performance of these approaches accords with intuition, as dropout is generally considered the manifestation of a process of disengagement, rather than a discrete decision (Balfanz, Herzog, and MacIver 2007; Programa Estado de la Nacion 2013; Frazelle and Nagel 2015). However, the high data requirements and relative complexity of these approaches limit their usefulness for most school systems.

The next best performing methods are in fact much more simple and usable. The ‘On-Track Indicator’ used in Chicago is a binary indicator of two underlying measures of students’ performance in the first grade of secondary school: earning the minimum number of course credits required for promotion to the next grade, and not earning more than one semester ‘F’ in any core courses. This indicator achieves 75% sensitivity and a 16% false-positive rate, and is in active use in the Chicago public school system (Allensworth and Easton 2005, 2007). Similarly, Bowers (2010) finds that logistic regressions including indicators of ever having repeated a grade and annual grade point average are able to achieve 81% specificity and a 25% false-positive rate in two school districts in the U.S.

While these results assess dropout prediction based on performance in-sample, the true test for the accuracy of prediction models is performance outside of the sample on which the models are estimated. Using administrative data on attendance, behavioral violations, standardized test scores, and demographic characteristics from the state of Wisconsin, Knowles (2015) estimates a range of models, from standard logistic regression to support vector machines, on a "training sample" and then tests their accuracy out of sample, finding that logistic regression models perform very well relative to more complex algorithms. Similarly, using rich administrative data in Denmark, Sara et al (2015) show that machine learning algorithms "trained" on one data set can achieve sensitivities over 90% (with false-positive rates around 20%) out of sample when predicting high school dropout over three-month intervals.

Overall, then, evidence is substantial that who will drop out can be accurately predicted using high-quality administrative data in the U.S. as well as Europe. We are not aware, however, of any studies investigating this question in the context of a developing country, where data requirements have only recently been met and dropout is often a much more widespread issue. This paper therefore provides one of the first applications of dropout prediction methodologies in lower middle-income education systems, made possible by the remarkable efforts of the ministries of education in both Guatemala and Honduras. Furthermore, unlike much of the existing U.S. research, we estimate models at the national level instead that at the state/province level. This is possible because in both countries the administrative data are harmonized, aggregated, and analyzed at the federal level.

3. Administrative data and dropouts in Guatemala and Honduras

3.1 Guatemala

Over the last several years, the Guatemalan Ministry of Education (*Ministerio de Educación* - MINEDUC) has substantially improved and expanded their administrative information system based on student-level records. One of the main innovations was the implementation of a unique student identifier, which started in upper secondary in 2009 and was progressively expanded until all students in primary and secondary schools were assigned unique identifiers in 2011. These identifiers are used by each school at the beginning of the school year to provide data on enrollment to the Ministry. Specifically, schools provide a list of all students who are enrolled in each grade, with their unique identifiers, and the Ministry centralizes and consolidates this information in a database that contains the annual enrollment status of all students (nearly 4 million each year) from 2011 to 2016. This data structure allows the educational trajectories of students to be followed through time, and specifically enables the identification of those who drop out and the year in which they left school. In addition to students' unique identifiers, the data also include information on student's gender, school year, and results of their October final examination (promoted, not promoted, withdrew), as well as school's sector (formal public, private, cooperative, municipal), teaching modality (bilingual, monolingual), municipality, department, and school identifier. [2,3]

These basic administrative microdata on enrollment can then be matched with additional data on student and school characteristics from the National Evaluation of Students. These assessments have been conducted in the first, third and sixth grades of primary education, the last grade of lower secondary education (ninth grade), and the last grade of upper secondary, with variable frequency.[4] The evaluations in secondary education are applied to all the students in the

[2] The formal public sector is the group of schools that are under the direct administration of MINEDUC. The municipal sector is composed of public schools administered by municipalities. The cooperative sector comprises public schools with a tripartite administration that involves the participation of MINEDUC, the municipality, and student's parents.

[3] Bilingual schools are institutions where instruction is offered in both Spanish and an indigenous language.

[4] The National Evaluations are conducted every year in the last grade of upper secondary, but the implementation in other grades has been less regular: 2006, 2009, and 2013 for ninth grade; 2006, 2007, 2008, 2009, 2010, 2013, and 2014 for third and sixth grade; and 2006, 2008, 2009 and 2010 for first grade. While this could limit the capacity of implementing an early warning system every year, the Government of Guatemala and the World Bank are working on replicating the analysis but using students' grades collected on a yearly basis.

corresponding grade, while the assessments in primary school only cover a random sample of the students in the formal public sector, excluding those children attending private, cooperative, and municipal schools.

In this paper, we combine information from the enrollment database and the National Evaluations conducted in sixth and ninth grades in 2013 to follow the educational trajectories of the cohort of students evaluated that year. Specifically, we focus on the random sample of 19,000 students in public schools assessed in sixth grade (6.5% of the sixth graders in the sector, 5.7% of all sixth graders) and the census of 196,000 students in all schools evaluated in ninth grade. As discussed later, these are critical grades for school dropout in both Guatemala and Honduras. National Evaluations provide data not only on students' reading and math abilities, but also on several self-reported characteristics including their motivation and interest in learning, parents' characteristics (e.g. education, occupation), and household resources. While these data are quite rich, records are only complete for about 60% of the random sample, as many students do not fully complete the questionnaires – in all the analysis presented, we drop all incomplete records.[5] Because the analysis is done at the national level, we complement these data with information at the department level on the gender-specific high school wage premium from Guatemala's national household survey, the *Encuesta de Condiciones de Vida* (ENCOVI) 2014, as an attempt to capture some basic characteristics of the local labor market that may affect the decision to drop out of school. Table A1 in the Appendix presents the definitions of all the variables that are used in the analysis, as well as some descriptive statistics for the cohort of students who were evaluated in sixth and ninth grades in 2013.

Dropout rates calculated using the administrative enrollment data confirm that a large proportion of youth leave school before graduating from secondary school in Guatemala. In particular, we focus on three different periods in which dropout is particularly high: the transition from primary to lower secondary education (*Ciclo Básico*), during lower secondary education, and the transition from lower to upper secondary education (*Ciclo Diversificado*). Figure 2 shows the dropout rates in each of these periods for the cohort of students under analysis. Fully one-quarter of the students

[5] In results that are not presented, multiple imputation techniques for the missing data are used, producing results that are highly consistent with those presented.

evaluated in the last grade of primary education (sixth grade) dropped out in the transition from primary to lower secondary, 18% of the remaining three-quarters (14% overall) left during lower secondary school, and 32% of the students who managed to reach the last year of lower secondary dropped out in the transition to upper secondary. Taken altogether, these figures imply a survival rate from sixth to tenth grade of 42%, meaning that the majority of sixth grade students drop out during this four-year period.[6] This pattern is consistent with the information that emerges from household surveys in Guatemala: for example, Adelman and Székely (2016) find that for the cohort reaching age 18 in 2012-2014, the percentage of youth enrolled in any grade falls from 80% to 47% from age 12 to 16.

While administrative enrollment data have several advantages for the study of school dropout, the records are still in an incipient phase of development and therefore subject to some error. In Guatemala, school principals (and teachers) manually enter information into a web-based information system, creating the possibility of typographical mistakes, incomplete data, and other human errors. Therefore, it is possible that some students who are not listed as enrolled in a particular year are not real dropouts (i.e. our data could overestimate the dropout rate). Some inconsistencies in the educational trajectories of the students have also been found (e.g., students skipping grades from one year to the next). To address this, a new data platform was implemented in 2015, with automatic flags to substantially reduce inconsistencies and improve the quality of data entered. Nevertheless, as we discussed above, when we compare the survival rates estimated with both administrative and household survey data we find similar rates, suggesting that the measurement error of dropout is relatively small in our data set.

3.2 Honduras

In Honduras, the *Secretaría de Educación* (SEDUC) began collecting administrative records in 2013. Student information is reported by teachers in both public and private institutions, who input their rosters into a web-based system. These records gather information for all formal sector students in Honduras and allow the tracking of the same children over time (about 1.5 million).

[6] As previously mentioned, the dropout rates in the transition from primary to lower secondary education and within lower secondary are based on a sample of students in the official sector, which represents 87% of all the students in sixth grade in 2013.

SEDUC estimates a coverage rate of 97%, which has improved in the two subsequent years for which data are available.

Given the recent implementation of the database, it contains limited information. Besides individual enrollment status, the data include some demographics (gender and age), attendance rates, whether the child attends a public or private institution, and current grade. Unlike in Guatemala, more detailed socioeconomic information for students or their household is unavailable. While recent data exist on exam scores, they cannot be matched to our sample because they use different individual identifiers, an issue that has only recently been addressed by SEDUC.

We therefore augment SEDUC's student data with municipal-level indicators from several sources. First, we add a set of school supply and quality indicators. The number of schools per municipality is included to measure educational supply at the extensive margin. These indicators are weighted by their target population using data from the National Statistics Institute. We first calculate the number of children in each municipality of primary (6-11) and secondary (12-18) age. Then, the number of primary and secondary schools is divided by the target population (in thousands). To measure intensive-margin educational supply, we take the number of available teaching positions in 2015 and weigh them by the target population in each municipality.[7] Quality indicators correspond to municipality-level infrastructure indices from the 2013 school census collected for the School Infrastructure Master Plan (*Plan Maestro de Infraestructura Educativa*). These indices are on a scale from zero to one, and we include a summary measure that is a weighted average of six dimensions: furnishings, basic services, natural disasters, social threats, hydro-sanitary issues, and physical infrastructure.[8]

Second, we add a set of municipal-level socioeconomic indicators. Honduras is classified as one of the most violent countries in the world (World Bank 2011), but there remains limited evidence

[7] Available teaching positions do not necessarily correspond to filled vacancies. For instance, a municipality may have 10 available positions, but only 8 employed teachers. The data do not allow us to make this distinction.

[8] Each index is comprised of multiple indicators (Secretary of Education, 2014). The furnishings index captures the number of desks and blackboards. The basic services index is an average of access to electricity, running water, and sewers. The natural disaster index measures how unlikely it is that a school is affected by floods, landslides strong winds, and earthquakes. The social threat index measures how difficult it is to access and consume alcohol, cigarettes, and illicit drugs. The hydro-sanitary index captures the state of sinks, toilets, and urinals. The physical infrastructure index evaluates how well the building, classrooms, and environments are suited for students to receive lessons.

on how crime correlates with schooling outcomes. Students are assigned the annual homicide rate for their municipality (per 100,000 population), taken from the Online Police Statistical System (*Sistema Estadístico Policial en Línea* - SEPOL).[9] We also include the following municipal-level indicators from the 2001 Population Census: unsatisfied basic needs (UBNs) poverty, the share of ethnic minority population, average years of education, average pre-school enrollment rates, average rates of child labor, the adolescent birth rate, and the share of households with migrants outside Honduras.

Dropout rates in Honduras vary across the school cycle. While only 8.8% of children drop out in primary school, this fraction increases substantially thereafter. Like Guatemala, dropout is particularly high in the transition from primary to lower secondary education (*Ciclo Básico*), during lower secondary education, and in the transition from lower to upper secondary education (*Ciclo Diversificado*). Figure 3 shows that 37.4% abandon school in the transition from primary to lower secondary, 18.4% drop out within lower secondary, and about 33.2% do not continue into upper secondary. Taken altogether, these figures imply a survival rate from sixth to tenth grade of 34%, meaning that the majority of sixth grade students drop out during this four-year period.

These rates are somewhat higher than household survey estimates but follow the same pattern. For the last available household survey in 2014, which collects data on children's grade progression, the dropout rate for the transition into lower secondary is 27%, 13% within lower secondary, and 21% during the transition from lower to upper secondary. Adelman and Székely (2016) also report similar trends when using age groups.

As with the Guatemalan data, this method of data collection has only been recently implemented and there remains room for improvement. Teacher reporting of student outcomes is potentially prone to input errors, which may result in measurement bias, although we are unable to determine their extent or direction. As noted in the previous paragraph, one potential outcome is that we may overestimate dropout rates. However, the administrative data present many advantages over household surveys, mainly sample size and nationwide coverage. Data collection is also constantly improving, due to efforts from USINIEH to ensure correct data input and quality control at

[9] Municipal-level homicide rates are publicly available at the SEPOL website: www.sepol.hn.

different stages. Furthermore, as long as the inflated dropout rates are driven by classical (random) measurement error, this would just lower the accuracy of our model. In that scenario, our estimates give us a lower bound of the accuracy levels of the model, and hence we would expect a better performance as the quality of the panel improves over time. Indeed, as we will show in the next section, the models work better in Guatemala, where the measurement error of dropout is much lower.

4. Who will drop out: Empirical predictions

4.1 Results

In our analysis, we focus on three periods during which both Guatemala and Honduras lose the majority of students to dropout: the transition between primary and lower secondary school, within lower secondary, and the transition from lower secondary to upper secondary. For ease of exposition, below we present the results of using data available in year t to predict who will drop out in the transition from primary to lower secondary in year t+1. Additional results on predicting dropouts within lower secondary and in the transition from lower to upper secondary, are presented in the Appendix.

We follow the method of Knowles (2015) and Stuit et al (2016) in estimating dropout models in three steps. These prediction models are based on the conceptual framing of dropout as a decision that can be affected by a broad range of underlying factors discussed in Section 1, and include all of the available covariates that could reasonably capture one of these factors. In this way, prediction modeling differs from other exercises that estimate the correlational or causal relationships between specific factors and dropout. First, we estimate linear probability models with dropout as a binary outcome, using the individual/household, school, and municipality/department-level covariates described in the previous section. However, for Guatemala, given the large number of variables available to include in the models we add a ‘zero step’: we combine the information from several highly correlated variables that measure similar characteristics and construct indices using a Principal Component Analysis (Pearson 1901; Hotelling 1933; Jolliffe 2002). These indices were obtained as component scores for the first principal component and enable us to estimate a parsimonious model with almost no loss in the accuracy of the prediction. [10] Table 2 shows the different specifications used in each country.[11]

Treating the estimated y-hats as dropout probabilities, we then construct a ROC curve by varying the cutoff point above which a student is identified as a predicted dropout. Each possible cutoff has a particular sensitivity and false-positive proportion associated to it. Therefore, in the third step we select the point that minimizes the distance from perfect prediction (the (0,1) point) and evaluate the models at this optimal cutoff.

Figure 4 shows the ROC curves for each model. The graphs show that even simple models work much better than random guesses (represented by the 45-degree line). In Honduras, model 1—which only includes basic demographic variables and a few municipal-level covariates— is noticeably above of the 45-degree line. For example, if we consider a 25% false-positive proportion, we could gain almost 40 percentage points of sensitivity by including these basic variables. Similar results are observed in Guatemala when we consider individual-level data (including test scores). As we include more information, not surprisingly, the models become more accurate. In particular, adding school fixed effects improves the performance of the models substantially.

Figure 5 tests whether the specification that uses all available variables is statistically different than a random guess. We estimate sensitivity and false positives at 100 different cutoffs by bootstrap to obtain a 95% confidence interval. In both Guatemala and Honduras, we find that these models are significantly better than a random guess.

Table 3 assesses the models in terms of sensitivity, false-positive proportion, and overall accuracy at the optimal cutoff value. In our preferred specification, we are able to correctly identify 80% of sixth grade students who will drop out in the transition to lower secondary, with a false-positive proportion close to 20%. Importantly, these accuracy levels are comparable to those observed in

[10] See Tables A1(a) and A1(b) in the Appendix and the notes below these tables for a description of the variables involved in their computation.

[11] Estimation results are presented in the Appendix.

developed countries. Furthermore, as will be discussed later, these models allow targeting mechanisms that are more accurate than other commonly used approaches.

4.2 Out-of-sample performance

While these models perform very well on the data on which they are constructed, we conduct two validation exercises to assess their performance out of sample. First, following the prediction literature, for each country we conduct a K-fold cross-validation, where the data is randomly divided into K "folds" or sub-samples (Knowles 2015). We present the results here for K=5 folds, but obtain similar results varying K between 3 and 8.[12] The prediction models are then estimated on four of the folds, and predictions made based on the models for each of the five folds. We then calculate sensitivity and false-positive proportions in sample (the four folds on which the model is estimated) and out of sample (the left-out fifth fold). This procedure, starting with the random division of the data into folds, is repeated 100 times to obtain the statistics presented in Table 4. While the models' out-of-sample performance is worse than in-sample, it remains reasonably good compared to other prediction models used in the U.S. context and significantly better than random guessing (Bowers, Sprott, and Taff 2013).

In Honduras, we can also perform what is arguably the best type of validation exercise for the model's out of sample prediction. Specifically, we estimate Model 3 in 2014 to predict the likelihood of dropping out in 2015. We use the optimal cutoff for 2013 and classify students at risk and not at risk. Then we check the efficiency of the early warning system. Results are shown in Figure 6. The method performs quite well out of sample, with a sensitivity rate of 76.3% and a false-positive proportion of 15.9%.

Finally, we consider the accuracy of the targeting facilitated by these models in relation to other potential targeting approaches through a simple simulation. In the scenario considered, a dropout prevention program for sixth graders has a fixed budget of either $1M, $2M, or $4M US in total and costs $200 US per student to implement, meaning that 5,000, 10,000, or 20,000 students can

[12] While no hard-and-fast rules exist around the optimal K, tradeoffs between bias and variance can help guide the choice, and K= 5 and K = 10 are commonly used (Kohavi 1995; Borra and Di Ciaccio 2010).

participate. In Guatemala, approximately 90,000 sixth graders drop out before enrolling in seventh grade, while in Honduras, approximately 40,000 do. How should students be targeted for this program in each country? In Table 5, we compare the early warning models to two other possible targeting approaches – targeting students in the poorest municipalities and targeting schools with the highest dropout rates. In both countries, the early warning models perform substantially better than the other options in identifying students who will eventually drop out, particularly for the smaller program sizes, and targeting students based on these models rather than targeting poor municipalities or high-dropout schools could reduce misallocation of resources by 30 to 80%.[13] However, school-level targeting also performs well and could be the most suitable targeting approach for interventions with substantial economies of scale.

5. Conclusions

Many developing countries, including Guatemala and Honduras, are approaching universal primary school enrollment and have also made substantial progress in expanding access to secondary school. However, dropout is a pressing concern for policy makers in both countries, as fewer than 50% of young people succeed in completing a full course of basic education. Policies and programs aimed at reducing dropout may be made more effective when education systems are able to identify with reasonable accuracy the students who are most likely to leave school early.

In both Guatemala and Honduras, substantial advances in the scope and reliability of education administrative data have created an opportunity to do just that, as illustrated in this paper. Using routinely collected administrative data and relatively simple analytical techniques, we show that early warning models can accurately predict which students will drop out, providing potentially actionable information to system and school leaders. The results immediately suggest two areas for further exploration: improving the early warning analysis and moving from early warning analysis to "early warning systems".

[13] The calculation on reduced misallocation is based on comparing the dollars wasted in Table 3. In Guatemala, due to the presence of missing data in the covariates used in the early warning system, the sample has been restricted to be exactly the same across the three methods to allocate resources.

Our analysis augments the most basic administrative data with additional data to improve the quality of prediction (in the case of Guatemala, with periodic national exam data for primary students; in the case of Honduras, with household survey and census data). This augmentation could affect the consistency of the prediction from year to year, as the availability of household survey and other data varies over time. It could also make the approach more difficult to replicate for practitioners within ministries of education, who may not be familiar with accessing and analyzing these other data sources. However, as mentioned above, in both countries more and more variables are being routinely collected over time, which will reduce the need to pull data from other sources and also could serve to improve the quality of prediction. In addition, as mentioned for Guatemala, a greater focus on obtaining complete data for each student would be critical to improving the quality and usefulness of the predictions.

While this paper has focused on the feasibility of making accurate and replicable predictions of who is most likely to drop out, this is only a first step towards developing an effective early warning system, and a natural question immediately arises – what to do with the data? Maintaining confidentiality, avoiding negative labeling of students, and identifying resources to act appropriately on the predictions are all among the potential important concerns. Moreover, the predictions in and of themselves do not identify the main factors that put any given student at risk of dropping out, information which is critical to providing effective interventions. A complete discussion about setting up an effective early warning system merits its own paper, but many lessons learned have emerged from the experiences of school districts in the U.S. These include clearly communicating the meaning of the predictions; defining roles at all levels in terms of who should receive what information and who is responsible for taking what actions; empowering local school officials to identify and implement relevant, customized prevention measures; and taking an iterative approach to facilitate learning from initial pilots (O'Cummings and Therriault 2015). These and other lessons learned from international experiences, combined with a strong knowledge of local contexts, show a path forward for responding effectively to the dropout crisis in Guatemala, Honduras, and many other countries.

References

Adelman, Melissa and Miguel Székely. 2016. "School Dropout in Central America: An Overview of Trends, Causes, Consequences, and Promising Interventions." World Bank Policy Research Working Paper 7561. Washington, D.C.: World Bank.

Allensworth, E. M., & J.Q. Easton. 2007. *What Matters for Staying On-track and Graduating in Chicago Public High Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year*. Chicago: Consortium on Chicago School Research.

Avitabile, Ciro and Rafael de Hoyos. 2014. "Heterogeneous Effects of Information about the Returns to Schooling on Student Learning: Evidence from a Randomized Control Trial in Mexico." World Bank Working Paper. Washington, D.C.: World Bank.

Balfanz, Robert, Liza Herzog, and Douglas MacIver. 2007. "Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-Grades Schools: Early Identification and Effective Interventions." *Educational Psychologist* 42(4): 223-235.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics* 3(2): 167-195.

Bassi, Marina, Matias Busso, and Juan Muñoz. 2016. "Is the Glass Half Empty or Half Full? School Enrollment, Graduation, and Dropout Rates in Latin America." *Economia* (forthcoming).

Becker, Gary. *Human Capital*. New York: Columbia University Press, 1964.

Behrman, Jere, Rafael de Hoyos, and Miguel Székely. 2015. "Out of School and Out of Work: A Conceptual Framework for Investigating "Ninis" in Latin America and the Caribbean." Washington, DC: World Bank. Background paper for the "Out of School, Out of Work" study.

Bentaouet-Kattan, R., and Székely, M. 2015. "Patterns, Consequences and Possible Causes of Dropout in Upper Secondary Education in Mexico." *Education Research International*, Volume 2015, Article ID 676472.

Borra, Simone and Agostino Di Ciaccio. 2010. "Measuring the Prediction Error: A Comparison of Cross-Validation, Bootstrap and Covariance Penalty Methods." *Computational Statistics & Data Analysis* 54(12): 2976-2989.

Bowers, Alex. 2010. "Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts." *Journal of Educational Research* 103(3): 191–207.

Bowers, Alex and Ryan Sprott. 2012. "Why Tenth Graders Fail to Finish High School: A Dropout Typology Latent Class Analysis." *The Journal of Education for Students Placed at Risk (JESPAR)* 17(3): 129-148.

Bowers, Alex, Ryan Sprott, and Sherry Taff. 2013. "Do We Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity." *The High School Journal* 96(2): 77-100.

Cardenas, Mauricio, Rafael de Hoyos, and Miguel Székely. 2015. "Out of School and Out of Work Youth in Latin America: a Persistent Problem in a Decade of Prosperity." *Economia* 16(1).

Estado de la Nación. 2013. *Estado de la Educación Costarricense*. San José, Costa Rica: Estado de la Nación.

European Commission. 2013. *Early Warning Systems in Europe: Practice, Methods, and Lessons*. Thematic Working Group on Early School Leaving. Brussels: European Commission.

Frazelle, Sarah and Aisling Nagel. 2015. *A Practitioner's Guide to Implementing Early Warning Systems*. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educaitonal Laboratory Northwest.

Hotelling, H. 1933. "Analysis of a Complex of Statistical Variables Into Principal Components." *Journal of Educational Psychology*, 24: 417-441 and 498-520.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *The Quarterly Journal of Economics* 125 (2): 515-548.

Jolliffe, I.T. 2002. *Principal Component Analysis*. New York: Springer-Verlag New York, Inc.

Fryer, Roland. 2013. "Information and Student Achievement: Evidence from a Cellular Phone Experiment." NBER Working Paper 19113.

Gleason, P., and M. Dynarski. 2002. "Do We Know Whom to Serve? Issues in Using Risk Factors to Identify Dropouts." *Journal of Education for Students Placed at Risk* 7(1): 25–41.

Heller, Sara, Anuj Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold Pollack. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." NBER Working Paper No.21178. Cambridge, MA: National Bureau of Economic Research (NBER).

Janosz, M., I. Archambault, J. Morizot, and L.S. Pagani. 2008. "School Engagement Trajectories and their Differential Predictive Relations." *Journal of Social Issues* 64(1): 21–40.

Knowles, Jared. 2015. "Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin." *Journal of Educational Data Mining* 7(3): 18-67.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Presented at the International Joint Conference on Artificial Intelligence.

Muthén, B. O. 2004. *Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data.* In D. Kaplan (Ed.), The Sage handbook of quantitative methodology for the social sciences (pp. 345–370). Thousand Oaks, CA: Sage Publications.

Nguyen, Trang. 2008. "Information, Role Models, and the Perceived Returns to Education." Unpublished manuscript.

O'Cummings, Mindee and Susan Therriault. 2015. "From Accountability to Prevention: Early Warning Systems Put Data to Work for Struggling Students." AIR Early Warning Systems in Education Program. Washington, D.C.: American Institutes for Research (AIR).

Oreopoulos, Philip and Kjell Salvanes. 2011. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25(1): 159-184.

Patrinos, Harry and George Psacharopoulos. 2004. "Returns to Investment in Education: a Further Update." *Education Economics* 12(2): 111-134.

Pearson, Karl. 1901. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine*, Series 6, 2(11): 559-572.

Rogosa, David. 2005. "Statistical Misunderstandings of the Properties of School Scores and School Accountability." *Yearbook of the National Society for the Study of Education* 104(2): 147-174.

Sara, Nicolae-Bogdan, Rasmus Halland, Christian Igel, and Stephen Alstrup. 2015. "High-School Dropout Prediction Using Machine Learning: A Danish Large-Scale Study." *The European Symposium on Artificial Neural Networks (ESANN) Proceedings 2015*.

Sheskin, David. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman & Hall / CRC Press, 2004.

Stuit, David, Mindee O'Cummings, Heather Norbury, Jessica Heppen, Sonica Dhillon, Jim Lindsay, and Bo Zhu. 2016. *Identifying Early Warning Indicators in Three Ohio School Districts*. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educaitonal Laboratory Midwest.

United Nations. 2015. *The Millennium Development Goals Report*. New York: United Nations.

World Bank. 2016. EdStats. Retrieved from: http://datatopics.worldbank.org/education/

Table 1. Event table

		Actual event	
		Drop out	Graduate
Prediction	Drop out	A (True positive)	B (False positive)
	Graduate	C (False negative)	D (True negative)

Sensitivity = A / (A+C)
Precision = A / (A+B)
Specificity = D / (B+D)
False-Positive Proportion = *1-Specificity* = B / (B+D)

Figure 1. ROC curve illustration

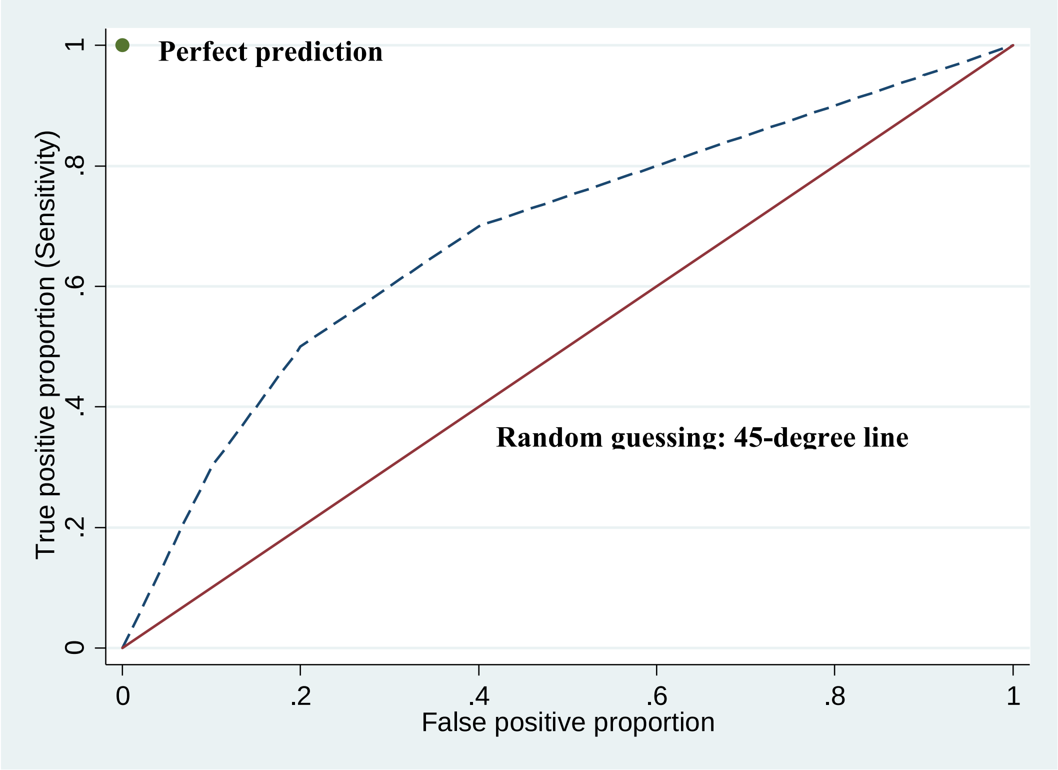
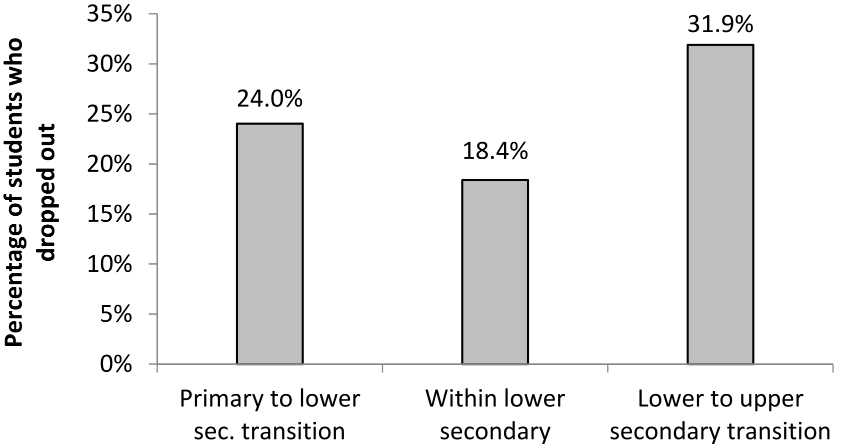


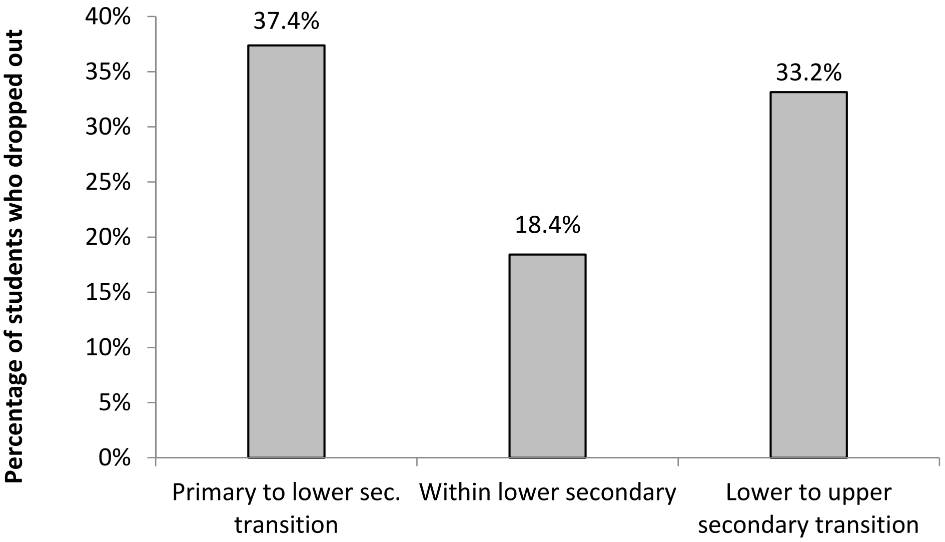
Figure 2. Dropout rates in Guatemala



Source: Author calculations based on administrative data on enrollment and the 2013 National Evaluations of students

Notes: (1) Dropout rates in primary to lower secondary transition computed as the percentage of sixth grade students evaluated in 2013 who did not enroll in 2014, 2015 and 2016. (2) Dropout rates within lower secondary computed as the percentage of sixth grade students evaluated in 2013 who were enrolled in 2014 but did not enroll in 2015 or 2016, excluding students who dropped out in the transition from primary to lower secondary from the calculation. (3) Dropout rates in lower to upper secondary transition computed as the percentage of ninth grade students evaluated in 2013 who did not enroll in 2014, 2015 and 2016.

Figure 3. Dropout rates in Honduras



Source: Author calculations based on USINIEH administrative data from 2013-2015

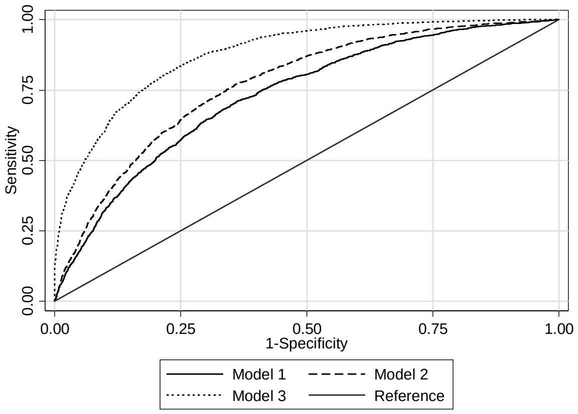
Notes: A dropout is a student who was enrolled in the previous year but did not enroll the following year.

Table 2. Covariates included in the models of dropout in the primary to lower secondary transition

Specification	Guatemala	Honduras
Model 1	**Individual level:** female, age, indigenous, works, housework, repeater, preschool, all_school_supplies, motivation, test_score	**Individual-level, supply-side, and municipal-level:** male, age, attendance rate, public school, rural school, number of secondary schools, number of teachers, summary infrastructure index, homicide rate, poverty (UBN), share of ethnic population, average years of education, average pre-school enrollment, average rate of child labor, adolescent birth rate, share of household with migrants
Model 2	**Individual and household level:** covariates in model 1, parents' maximum educational attainment (dummies), help_homework, low_quality_housing, electricity, pc, books, goods_availability_index	**Individual-level and municipality fixed effects:** male, age, attendance rate, public school, rural school, municipality fixed effects
Model 3	**Individual, household, and school/area of residence level:** covariates in model 2, school fixed effects, hs_wage_premium	**Individual-level and school fixed effects:** male, age, attendance rate, school fixed effects

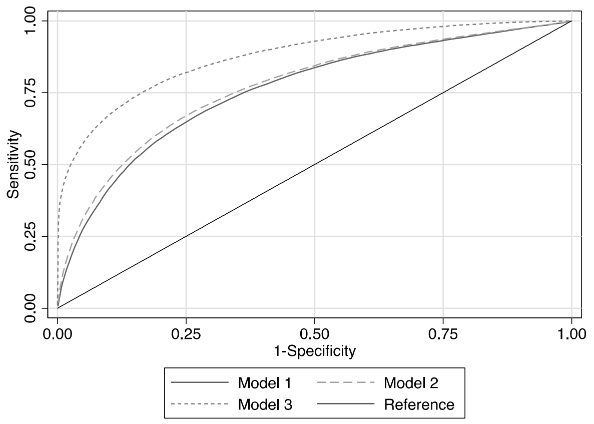
Figure 4. ROC curves of the different models of dropout in the primary to lower secondary transition

Guatemala



Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014).
Notes: Model 1: model with covariates at the individual level. Model 2: model with covariates at the individual and household level. Model 3: model with covariates at the individual, household, and school/area of residence level.

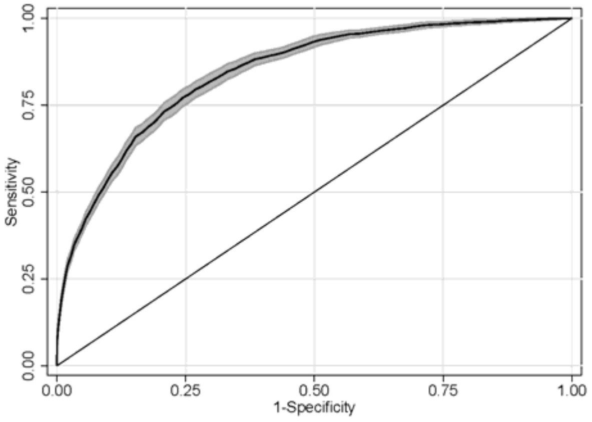
Honduras



Source: Author calculations based on USINIEH administrative data from 2013-2015.
Notes: Model 1: model with covariates at the individual, supply-side and municipal-level variables. Model 2: model with covariates at the individual and school-level, with municipality fixed effects. Model 3: model with covariates at the individual-level with school fixed effects.

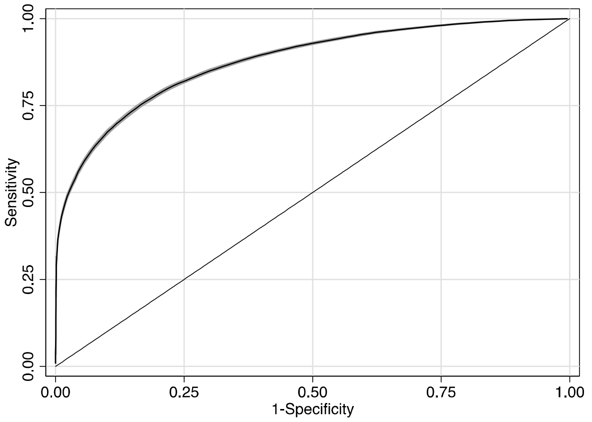
Figure 5. ROC curves for full models of dropout in the primary to lower secondary transition: bootstrap confidence intervals

Guatemala



Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014).
Notes: ROC curve for Model 3. Confidence intervals are calculated by 100 block bootstrap repetitions that resample students from departments.

Honduras



Source: Author calculations based on USINIEH administrative data from 2013-2015.
Notes: ROC curve for Model 3. Confidence intervals are calculated by 100 block bootstrap repetitions that resample students from schools.

Table 3. Performance of the models of dropout in the primary to lower secondary transition

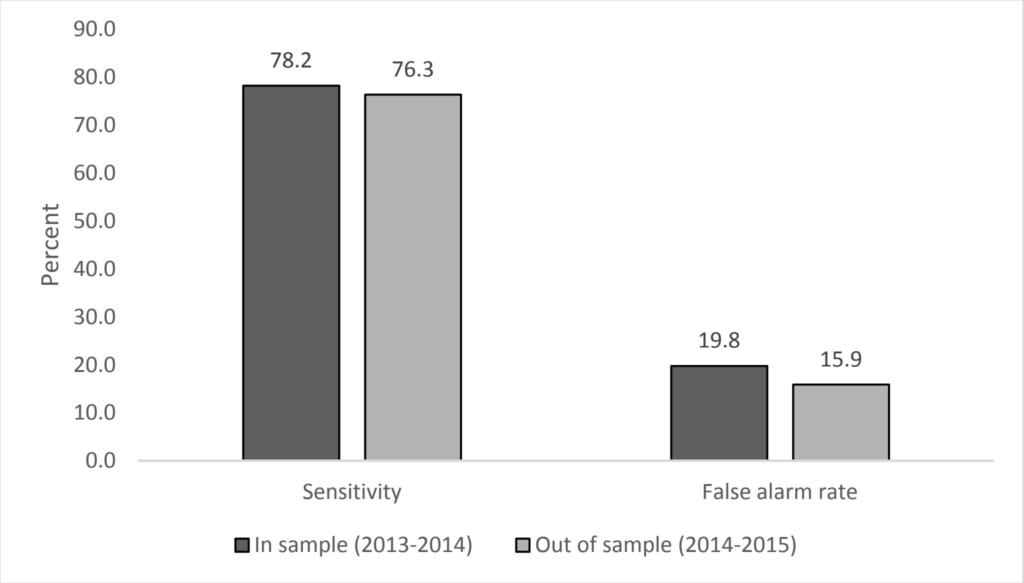
Country and specification	**Sensitivity (percent)**	**False-positive proportion (percent)**	**Overall accuracy[1] (proportion)**
Guatemala			
Model 1	67.8	32.9	0.67
Model 2	71.6	30.7	0.70
Model 3	80.0	21.2	0.79
Honduras			
Model 1	68.8	28.9	0.70
Model 2	69.1	27.2	0.71
Model 3	78.2	19.8	0.79

Source: Author calculations on administrative data from Guatemala and Honduras.
Note: This is the area-under-the-curve statistic, which ranges from .50 to 1.00, with higher values associated with higher accuracy (higher sensitivity and lower false-positive proportion).

Table 4. Out-of-sample performance in Guatemala and Honduras based on K-fold cross validation

	Mean	SD	Min	Max
Transition from primary to lower secondary: **Guatemala**				
Sensitivity				
In sample	81.27	0.52	80.10	82.40
Out of sample	69.26	0.72	67.95	70.67
False-Positive Proportion				
In sample	20.94	0.47	19.90	21.98
Out of sample	24.51	0.52	23.54	25.74
Transition from primary to lower secondary: **Honduras**				
Sensitivity				
In sample	78.26	0.16	77.94	78.64
Out of sample	72.27	0.19	71.87	72.73
False-Positive Proportion				
In sample	19.04	0.18	18.68	19.48
Out of sample	21.73	0.19	21.37	22.17

Figure 6. Out-of-sample performance of early warning system in Honduras



Source: Author calculations based on USINIEH administrative data from 2013-2015

Table 5. Relative performance of the early warning system as a method to allocate limited resources ($200 US per student)

Country	**Number of students**	**Method of allocation**								
		Municipality poverty rates			**School dropout rates**			**EWS**		
		True positives	*False positives*	*Dollars wasted*	*True positives*	*False positives*	*Dollars wasted*	*True positives*	*False positives*	*Dollars wasted*
Guatemala	5,000	45%	55%	$ 552,022	93%	7%	$ 72,772	99%	1%	$ 8,319
	10,000	43%	57%	$ 1,143,022	86%	14%	$ 277,037	89%	11%	$ 210,860
	20,000	35%	65%	$ 2,609,953	73%	27%	$ 1,086,566	78%	22%	$ 898,115
Honduras	5,000	58%	42%	$ 418,000	87%	13%	$ 129,000	99%	1%	$ 6,200
	10,000	60%	40%	$ 798,200	82%	18%	$ 364,000	99%	1%	$ 12,400
	20,000	56%	44%	$ 1,748,000	74%	26%	$ 1,046,800	95%	5%	$ 188,200

Source: Author calculations based on administrative data from Guatemala and Honduras.
Notes: The allocation of resources by municipality poverty rates selects the students who live in the municipalities with the highest poverty rates. The allocation of resources by school dropout rates selects the students who are in the schools with highest dropout rates in the sixth to seventh grade transition. The allocation of resources using the early warning system selects the students who have the highest predicted probability of dropout according to the model in Table A6. True positives are the percentage of hypothetical beneficiaries who dropout in the transition from primary to lower secondary, and false positives are the percentage who do not dropout in this transition. Dollars wasted are computed as $200 x Number of students x False positives.

Appendix

Table A1. Variable definitions and summary statistics, Guatemala

(a) Guatemala: cohort of students evaluated in sixth grade in 2013

Name of variable	*Definition*	*Observations*	*Mean*	*Std. Dev.*
female	1 = Female	18197	0.50	0.50
age	Age (in years)	18176	13.02	1.17
indigenous	1 = Indigenous (speaks maya, xinka or garifuna)	18196	0.20	0.40
works	1 = works	17591	0.34	0.47
housework	1 = helps with housework	17725	0.96	0.20
repeater	1 = Repeated a grade	17918	0.30	0.46
preschool	1 = Attended preprimary education	17643	0.73	0.44
all_school_supplies	1 = Student has all the school supplies (books, exercise books, pencil, etc.)	18197	0.56	0.50
motivated_bylearning	1 = Student attends school because she is motivated by learning	18099	0.68	0.47
likes_reading	1 = Student likes reading	16385	0.97	0.17
motivation	Index of motivation at school [1]	16309	0.00	1.00
reading_score	Measure of student's ability in reading (expressed in logits)	18147	0.06	0.94
math_score	Measure of student's ability in math (expressed in logits)	17908	0.17	0.83
test_score	Index of reading and math test scores [1]	17872	0.00	1.00
noschool	1 = Parents did not attend school	15828	0.26	0.44
primary	1 = Parents' maximum educational attainment is primary	15828	0.44	0.50
lowersecondary	1 = Parents' maximum educational attainment is lower secondary	15828	0.14	0.35
uppersecondary	1 = Parents' maximum educational attainment is upper secondary	15828	0.06	0.23
some_college	1 = Parents' maximum educational attainment is some college	15828	0.03	0.17
college	1 = Parents' maximum educational attainment is complete college	15828	0.07	0.26
help_homework	1 = Someone at home helps the student with homework	17531	0.43	0.50
low_quality_floor	1 = Floor made of low quality materials (soil)	18016	0.25	0.43
low_quality_walls	1 = Walls made of low quality materials (sheet)	18061	0.06	0.24
low_quality_water	1 = Water source is of low quality (natural source or well)	17957	0.29	0.45
low_quality_housing	Index of low quality housing [1]	17684	0.00	1.00
electricity	1 = Electricity at home	17967	0.85	0.36
pc	1 = Personal computer at home	18120	0.31	0.46
books	1 = Presence of books at home for reading	16723	0.92	0.26
car	1 = Car at home	18120	0.26	0.44
cell_phone	1 = Cell phone at home	18120	0.83	0.38
tv	1 = tv at home	18120	0.78	0.42
refrigerator	1 = Refrigerator at home	18120	0.43	0.49
sound_system	1 = Sound system at home	18120	0.52	0.50
dvd_vhs	1 = Dvd/VHS at home	18120	0.57	0.50
washer	1 = Washer at home	18120	0.17	0.38
gas_stove	1 = Gas stove at home	18120	0.36	0.48
iron	1 = Iron at home	18120	0.56	0.50
truck	1 = Truck/tractor at home	18120	0.03	0.17
goods_availability_index	Index of goods availability [1]	18123	0.00	1.00
hs_wage_premium	Gender specific high-school wage premium in the department (percentage points)	18197	120.40	80.09

Note: 1. Indices are computed combining the information of several variables by using a principal component analysis. These variables are: motivated_bylearning and likes_reading for motivation; reading_score and math_score for test_score, low_quality_floor, low_quality_walls, and low_quality_water for low_quality_housing; car, cell_phone, tv, refrigerator, sound_system, dvd_vhs, washer, gas_stove, iron, and truck for goods_availability_index.

(b) Guatemala: cohort of students evaluated in ninth grade in 2013

Name of variable	*Description*	*Observations*	*Mean*	*Std. Dev.*
female	1 = Female	196516	0.48	0.50
age	Age (in years)	196334	16.50	3.64
indigenous	1 = Indigenous (mother tongue is maya, xinka or garifuna)	195992	0.18	0.38
preschool	1 = Attended preprimary education	194167	0.84	0.37
repeater	1 = Repeated a grade in primary	192498	0.34	0.47
more1hour_to_school	1 = Time from home to school: more than one hour	195652	0.05	0.22
works	1 = works	191315	0.33	0.47
read_newspaper	1 = Reads the newspaper	195733	0.96	0.19
read_books	1 = Reads books for pleasure or personal interest	195040	0.80	0.40
motivation	Index of motivation at school [1]	194612	0.00	1.00
reading_score	Measure of student's ability in reading (expressed in logits)	195726	-0.46	0.78
math_score	Measure of student's ability in math (expressed in logits)	196011	0.00	0.66
test_score	Index of reading and math test scores [1]	195608	0.00	1.00
noschool	1 = Parents did not attend school	193846	0.12	0.33
primary	1 = Parents' maximum educational attainment is primary	193846	0.48	0.50
lowersecondary	1 = Parents' maximum educational attainment is lower secondary	193846	0.14	0.35
uppersecondary	1 = Parents' maximum educational attainment is upper secondary	193846	0.14	0.34
college	1 = Parents' maximum educational attainment is complete college	193846	0.09	0.28
postgraduate	1 = Parents' maximum educational attainment is posgraduate education	193846	0.02	0.15
low_quality_floor	1 = Floor made of low quality materials (soil)	195967	0.15	0.36
low_quality_walls	1 = Walls made of low quality materials (sheet)	196074	0.05	0.21
low_quality_roof	1 = Roof made of low quality materials (fragile or perishable)	195886	0.01	0.08
low_quality_water	1 = Water source is of low quality (natural source or well)	194945	0.17	0.38
low_quality_housing	Index of low quality housing [1]	193683	0.00	1.00
electricity	1 = Electricity at home	195764	0.95	0.22
pc	1 = Personal computer at home	187903	0.49	0.50
car	1 = Car at home	184132	0.43	0.49
cell_phone	1 = Cell phone at home	195906	0.94	0.24
tv	1 = tv at home	187903	0.94	0.23
refrigerator	1 = Refrigerator at home	187903	0.65	0.48
sound_system	1 = Sound system at home	187903	0.69	0.46
dvd_vhs	1 = Dvd/VHS at home	187903	0.59	0.49
washer	1 = Washer at home	187903	0.25	0.43
goods_availability_index	Index of goods availability [1]	176146	0.00	1.00
hs_wage_premium	Gender specific high-school wage premium in the local labor market	196511	95.96	74.03

Note: 1. Indices are computed combining the information of several variables by using a principal component analysis. These variables are: read_newspaper and read_books for motivation; reading_score and math_score for test_score, low_quality_floor, low_quality_walls, low_quality_roof, and low_quality_water for low_quality_housing; car, cell_phone, tv, refrigerator, sound_system, dvd_vhs, and washer for goods_availability_index.

Table A2. Variable definitions and summary statistics, Honduras

Name of variable	Definition	Observations	Mean	Std. Dev.
male	=1 if male	613,973	0.49	0.50
age	Age (in years)	613,973	14.09	1.58
attendance rate	Attendance rate	613,973	1.00	0.02
public school	=1 if attends a public school	445,777	0.78	0.42
rural school	=1 if the school is classified as rural	445,777	0.31	0.46
schools	Number of secondary schools per 1,000 secondary-age children in municipality	445,777	5.84	3.22
teachers	Number of teaching positions per 1,000 school-age children in municipality	445,777	36.46	19.56
summary infrastructure index	Average school infrastructure index by municipality	445,777	0.53	0.06
homicide rate	Municipality homicide rate (per 100,000 people)	445,777	70.63	51.96
poverty	Unsatisfied basic needs poverty rate (2001 census)	613,973	53.64	15.88
ethnic	Share of Ethnic Population (2001 census)	613,973	0.06	0.15
years of education	Average Years of Education (2001 census)	613,973	6.38	1.48
pre-school enrollment	Average Pre-school Enrollment (2001 census)	613,973	0.34	0.12
child labor	Average Child Labor Rate (2001 census)	613,973	0.12	0.04
adolescent birthrate	Average Adolescent Birth Rate (2001 census)	613,973	92.90	16.00
migrants	Average Share of Household with Migrants (2001 census)	613,973	0.03	0.02

Table A3. Model of dropout in the primary to lower secondary transition in Guatemala.

	Model		
	(1)	(2)	(3)
Individual level variables:			
female	0.064***	0.057***	0.068***
	(0.008)	(0.008)	(0.009)
age	0.056***	0.042***	0.044***
	(0.004)	(0.004)	(0.004)
indigenous	0.116***	0.068***	-0.022
	(0.018)	(0.018)	(0.020)
works	0.075***	0.053***	0.042***
	(0.011)	(0.012)	(0.011)
housework	-0.009	-0.020	-0.028
	(0.019)	(0.022)	(0.023)
repeater	-0.019*	-0.020*	-0.011
	(0.010)	(0.011)	(0.010)
preschool	-0.085***	-0.056***	-0.051***
	(0.010)	(0.011)	(0.011)
all_school_supplies	-0.010	0.009	-0.002
	(0.008)	(0.009)	(0.008)
motivation	-0.013***	-0.012***	-0.012***
	(0.004)	(0.004)	(0.004)
test_score	-0.063***	-0.042***	-0.023***
	(0.005)	(0.005)	(0.005)
Household level variables:			
primary		-0.086***	-0.050***
		(0.013)	(0.011)
lowersecondary		-0.142***	-0.067***
		(0.015)	(0.014)
uppersecondary		-0.130***	-0.068***
		(0.017)	(0.016)
some_college		-0.108***	-0.043**
		(0.021)	(0.021)
college		-0.122***	-0.056***
		(0.016)	(0.015)
help_homework		-0.034***	-0.000
		(0.008)	(0.008)
low_quality_housing		0.021***	0.023***
		(0.006)	(0.005)
electricity		-0.028	-0.041**
		(0.018)	(0.017)
pc		-0.054***	-0.041***
		(0.008)	(0.009)
books		-0.022	0.001
		(0.017)	(0.017)
goods_availability_index		-0.035***	-0.019***
		(0.006)	(0.006)
School/area of residence variables:			
hs_wage_premium			0.000
			(0.000)
Constant	-0.490***	-0.156**	-0.367***
	(0.057)	(0.066)	(0.062)
Observations	13,918	10,785	10,785
R-squared	0.115	0.155	0.363

Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014).

Notes: (1) Standard errors clustered by department in parentheses. (2) *** p<0.01, ** p<0.05, * p<0.1 (3) Dependent variable: 1 = dropout in the sixth to seventh grade transition. (4) School/area of residence variables include school fixed effects.

Table A4. Model of dropout in the primary to lower secondary transition in Honduras.

	Model		
	(1)	(2)	(3)
Male	-0.000	-0.000	0.004
	(0.003)	(0.003)	(0.003)
Age	0.098	0.099	0.086
	(0.003)***	(0.003)***	(0.002)***
Attendance rate	-0.887	-1.013	-1.477
	(0.153)***	(0.150)***	(0.117)***
Public School	-0.018	-0.019	
	(0.014)	(0.014)	
Rural school	0.220	0.236	
	(0.019)***	(0.022)***	
Number of Schools	0.005		
	(0.010)		
Number of Teachers	-0.001		
	(0.001)		
School Infrastructure Index	0.209		
	(0.132)		
Homicide rate	0.000		
	(0.000)**		
UBN Poverty Rate	0.003		
	(0.001)***		
Average share of Ethnic Population (2001)	-0.084		
	(0.044)*		
Average Years of Education (2001)	-0.041		
	(0.014)***		
Average Pre-school Enrollment (2001)	0.108		
	(0.062)*		
Average Child Labor Rate (2001)	0.927		
	(0.130)***		
Average Adolescent Birth Rate (2001)	-0.000		
	(0.000)		
Average Share of Household with Migrants (2001)	0.566		
	(0.253)**		
Constant	-0.224	0.092	0.809
	(0.184)	(0.132)	(0.120)***
Grade Fixed Effects	Yes	Yes	Yes
Population Polynomial	Cubic	No	No
R^2	0.193	0.214	0.379
Observations	109,226	109,226	109,226

Source: Author calculations based on USINIEH administrative data from 2013-2015. Notes: Clustered Standard Errors by Municipality in Parentheses

Significant at *10%, **5%, ***1%

Figure A1. ROC curves of the different models of dropout within lower secondary.

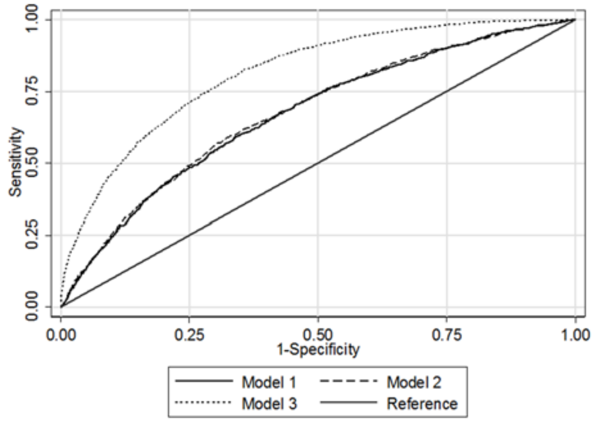
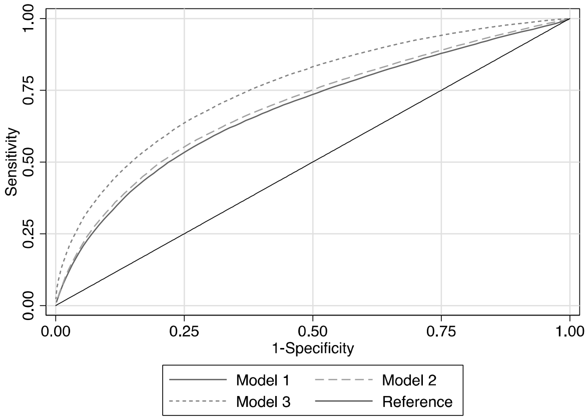
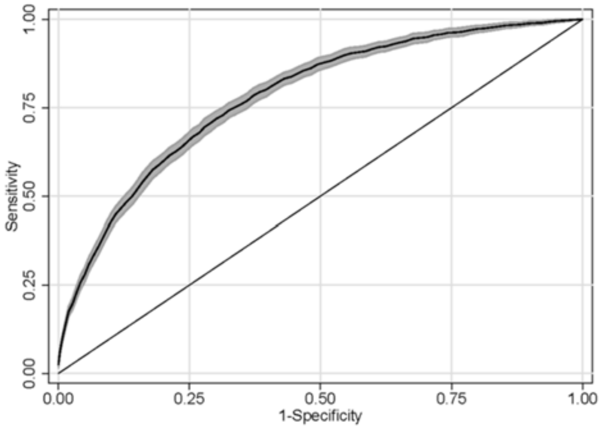
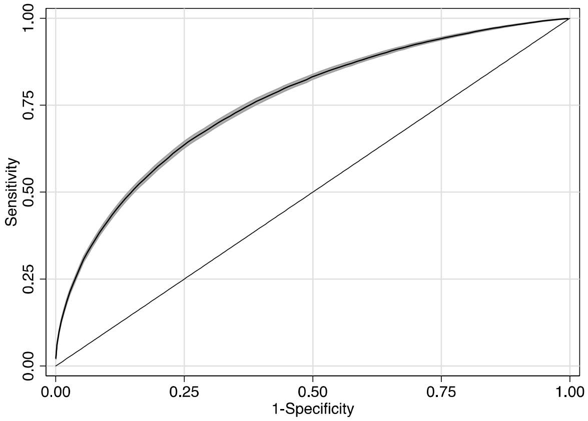
Guatemala	Honduras
	
	
Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014). *Notes*: Model 1: model with covariates at the individual level. Model 2: model with covariates at the individual and household level. Model 3: model with covariates at the individual, household, and school/area of residence level.	*Source*: Author calculations based on USINIEH administrative data from 2013-2015. *Notes*: Model 1: model with covariates at the individual, supply-side and municipal-level variables. Model 2: model with covariates at the individual and school-level, with municipality fixed effects. Model 3: model with covariates at the individual-level with school fixed effects.

Table A5. Performance of the models of dropout within lower secondary.

Country and specification	Sensitivity (percent)	False-positive proportion (percent)	Overall accuracy[1] (proportion)
Guatemala			
Model 1	58.2	33.8	0.62
Model 2	61.6	35.3	0.63
Model 3	74.2	27.4	0.73
Honduras			
Model 1	60.5	32.1	0.64
Model 2	62.8	33.0	0.65
Model 3	68.1	29.6	0.69

Source: Author calculations based on administrative data from Guatemala and Honduras.

Notes: 1. The area-under-the-curve statistic, which ranges from .50 to 1.00, with higher values associated with higher accuracy (higher sensitivity and lower false-positive proportion).

Table A6. Model of dropout within lower secondary in Guatemala.

	Model		
	(1)	(2)	(3)
Individual level variables:			
female	0.005	0.014	0.009
	(0.008)	(0.009)	(0.011)
age	0.056***	0.058***	0.065***
	(0.005)	(0.005)	(0.006)
indigenous	0.025**	0.019	-0.032
	(0.013)	(0.014)	(0.023)
works	0.015	0.010	0.014
	(0.009)	(0.011)	(0.013)
housework	0.004	0.017	0.024
	(0.019)	(0.022)	(0.025)
repeater	0.027***	0.028**	0.019
	(0.010)	(0.011)	(0.013)
preschool	-0.048***	-0.038***	-0.028**
	(0.010)	(0.011)	(0.013)
all_school_supplies	-0.000	-0.006	-0.003
	(0.008)	(0.009)	(0.011)
motivation	-0.004	-0.002	-0.001
	(0.004)	(0.004)	(0.005)
test_score	-0.022***	-0.020***	-0.022***
	(0.004)	(0.005)	(0.006)
Household level variables:			
primary		-0.022*	-0.018
		(0.013)	(0.014)
lowersecondary		-0.022	-0.015
		(0.016)	(0.017)
uppersecondary		-0.035*	-0.025
		(0.018)	(0.021)
some_college		-0.023	-0.007
		(0.023)	(0.026)
college		-0.026	-0.019
		(0.017)	(0.019)
help_homework		-0.018**	-0.019**
		(0.008)	(0.010)
low_quality_housing		0.006	0.009
		(0.006)	(0.006)
electricity		0.029*	0.020
		(0.017)	(0.020)
pc		-0.021**	-0.019*
		(0.010)	(0.011)
books		-0.013	-0.019
		(0.017)	(0.019)
goods_availability_index		0.007	0.005
		(0.006)	(0.007)
School/area of residence variables:			
hs_wage_premium			-0.000
			(0.000)
Constant	-0.523***	-0.554***	-0.539***
	(0.061)	(0.075)	(0.085)
Observations	10,879	8,375	8,375
R-squared	0.048	0.054	0.205

Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014).

Notes: (1) Standard errors clustered by department in parentheses. (2) *** p<0.01, ** p<0.05, * p<0.1 (3) Dependent variable: 1 = dropout within lower secondary. (4) School/area of residence variables include school fixed effects. (5) Students who dropped out in the transition from primary to lower secondary are excluded from the sample.

Table A7. Model of dropout within lower secondary in Honduras.

	Model		
	(1)	(2)	(3)
Male	0.012	0.012	0.013
	(0.003)***	(0.003)***	(0.002)***
Age	0.078	0.077	0.074
	(0.002)***	(0.002)***	(0.002)***
Attendance rate	-0.323	-0.347	-0.334
	(0.107)***	(0.111)***	(0.115)***
Public School	0.045	0.049	
	(0.010)***	(0.011)***	
Rural school	-0.017	-0.022	
	(0.009)**	(0.010)**	
Number of Schools	0.022		
	(0.005)***		
Number of Teachers	-0.001		
	(0.000)***		
School Infrastructure Index	-0.019		
	(0.069)		
Homicide rate	0.000		
	(0.000)		
UBN Poverty Rate	0.001		
	(0.000)*		
Average share of Ethnic Population (2001)	-0.011		
	(0.021)		
Average Years of Education (2001)	0.002		
	(0.007)		
Average Pre-school Enrollment (2001)	0.009		
	(0.033)		
Average Child Labor Rate (2001)	-0.090		
	(0.080)		
Average Adolescent Birth Rate (2001)	0.000		
	(0.000)		
Average Share of Household with Migrants (2001)	0.606		
	(0.146)***		
Constant	-0.665	-0.527	-0.465
	(0.151)***	(0.129)***	(0.123)***
Grade Fixed Effects	Yes	Yes	Yes
Population Polynomial	Cubic	No	No
R^2	0.077	0.086	0.153
Observations	156,842	156,842	156,842

Source: Author calculations based on USINIEH administrative data from 2013-2015.

Notes: Clustered Standard Errors by Municipality in Parentheses
Significant at *10%, **5%, ***1%

Figure A2. ROC curves of the different models of dropout in the lower to upper secondary transition.

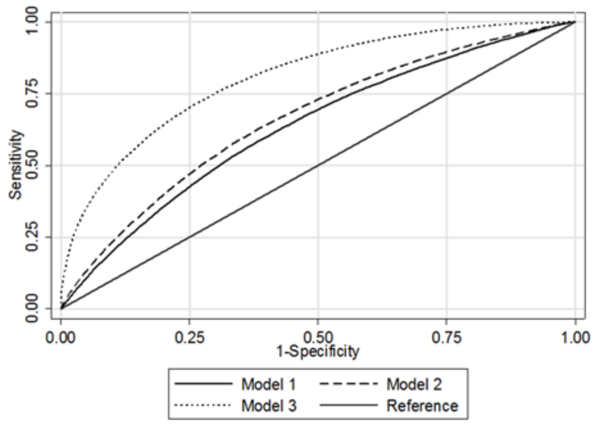
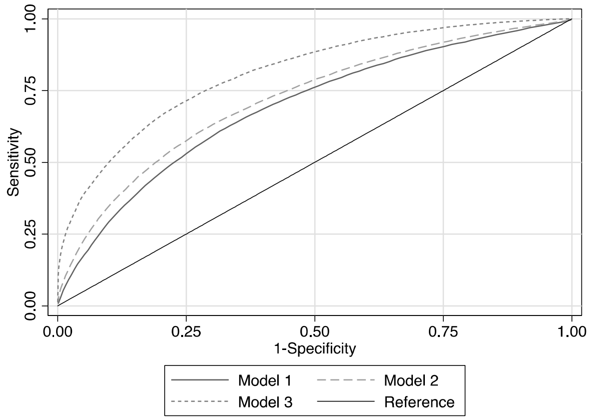
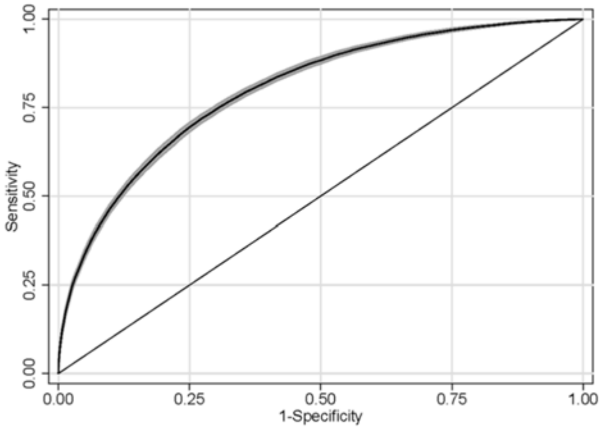
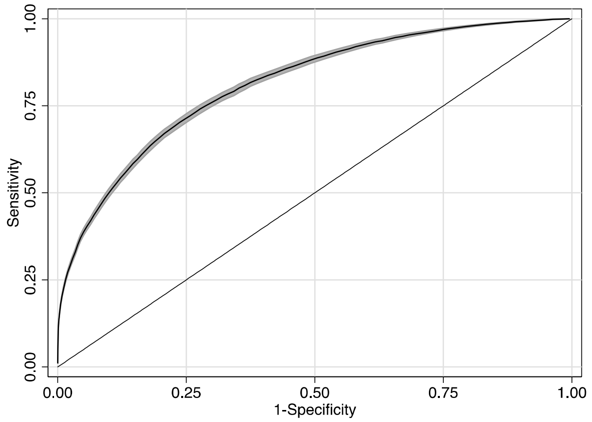
Guatemala	Honduras
	
	
Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014). *Notes*: Model 1: model with covariates at the individual level. Model 2: model with covariates at the individual and household level. Model 3: model with covariates at the individual, household, and school/area of residence level.	*Source*: Author calculations based on USINIEH administrative data from 2013-2015. *Notes*: Model 1: model with covariates at the individual, supply-side and municipal-level variables. Model 2: model with covariates at the individual and school-level, with municipality fixed effects. Model 3: model with covariates at the individual-level with school fixed effects.

Table A8. Performance of the models of dropout in the lower to upper secondary transition.

Country and specification	Sensitivity (percent)	False-positive proportion (percent)	Overall accuracy[1] (proportion)
Guatemala			
Model 1	60.2	39.0	0.61
Model 2	62.6	38.3	0.62
Model 3	72.9	27.7	0.73
Honduras			
Model 1	62.9	33.6	0.65
Model 2	64.8	31.7	0.67
Model 3	73.5	27.0	0.73

Source: Author calculations based on administrative data from Guatemala and Honduras.

Notes: 1. The area-under-the-curve statistic, which ranges from .50 to 1.00, with higher values associated with higher accuracy (higher sensitivity and lower false-positive proportion).

Table A9. Model of dropout in the lower to upper secondary transition in Guatemala.

	Model		
	(1)	(2)	(3)
Individual level variables:			
female	-0.015***	-0.014***	-0.008***
	(0.004)	(0.004)	(0.003)
age	0.003	0.002	0.004*
	(0.002)	(0.002)	(0.003)
indigenous	0.120***	0.062***	0.018***
	(0.008)	(0.008)	(0.006)
preschool	-0.037***	-0.002	0.002
	(0.004)	(0.004)	(0.004)
repeater	0.046***	0.033***	0.029***
	(0.004)	(0.004)	(0.003)
more1hour_to_school	0.031***	0.025***	0.032***
	(0.011)	(0.009)	(0.007)
works	0.082***	0.054***	0.028***
	(0.013)	(0.011)	(0.003)
motivation	0.000	-0.002*	-0.001
	(0.001)	(0.001)	(0.001)
test_score	-0.050***	-0.021***	-0.009***
	(0.002)	(0.002)	(0.001)
Household level variables:			
primary		-0.037***	-0.030***
		(0.006)	(0.005)
lowersecondary		-0.096***	-0.058***
		(0.008)	(0.005)
uppersecondary		-0.124***	-0.068***
		(0.010)	(0.005)
college		-0.128***	-0.058***
		(0.010)	(0.006)
postgraduate		-0.143***	-0.058***
		(0.012)	(0.008)
low_quality_housing		0.024***	0.012***
		(0.002)	(0.002)
electricity		-0.102***	-0.055***
		(0.011)	(0.009)
pc		-0.052***	-0.034***
		(0.005)	(0.003)
goods_availability_index		-0.010***	-0.002
		(0.002)	(0.002)
School/area of residence variables:			
hs_wage_premium			-0.000
			(0.000)
Constant	0.249***	0.439***	0.171***
	(0.026)	(0.027)	(0.038)
Observations	181,225	160,441	160,441
R-squared	0.052	0.067	0.263

Source: Author calculations based on administrative data on enrollment, National Evaluations of students (2013), and ENCOVI (2014).

Notes: (1) Standard errors clustered by department in parentheses. (2) *** p<0.01, ** p<0.05, * p<0.1 (3) Dependent variable: 1 = dropout in the lower to upper secondary transition. (4) School/area of residence variables include school fixed effects.

Table A10. Model of dropout in the lower to upper secondary transition in Honduras.

	Model		
	(1)	(2)	(3)
Male	0.023	0.023	0.021
	(0.005)***	(0.005)***	(0.004)***
Age	0.092	0.090	0.082
	(0.004)***	(0.004)***	(0.003)***
Attendance rate	-0.114	-0.031	-0.135
	(0.143)	(0.166)	(0.096)
Public School	0.054	0.068	
	(0.018)***	(0.016)***	
Rural school	0.166	0.175	
	(0.017)***	(0.020)***	
Number of Schools	0.006		
	(0.011)		
Number of Teachers	-0.001		
	(0.001)		
School Infrastructure Index	0.076		
	(0.174)		
Homicide rate	-0.000		
	(0.000)		
UBN Poverty Rate	-0.000		
	(0.001)		
Average share of Ethnic Population (2001)	0.157		
	(0.046)***		
Average Years of Education (2001)	-0.042		
	(0.015)***		
Average Pre-school Enrollment (2001)	0.031		
	(0.075)		
Average Child Labor Rate (2001)	0.128		
	(0.170)		
Average Adolescent Birth Rate (2001)	-0.001		
	(0.000)***		
Average Share of Household with Migrants (2001)	-0.411		
	(0.328)		
Constant	-0.715	-1.073	-0.753
	(0.237)***	(0.165)***	(0.106)***
Grade Fixed Effects	Yes	Yes	Yes
Population Polynomial	Cubic	No	No
R^2	0.099	0.133	0.256
Observations	53,142	53,142	53,142

Source: Author calculations based on USINIEH administrative data from 2013-2015.

Notes: Clustered Standard Errors by Municipality in Parentheses
Significant at *10%, **5%, ***1%