An Empirical Model of Educational Attainment in Uganda's Youth Population

Yvonne Dixon
TWSB:
Economic Analyst

I. INTRODUCTION

Recent progress in economic growth theory has revealed the significance of human capital in economic growth and development. The contributions of Mankiw et al. (1992) provided the first insights into the importance of human capital in macroeconomic growth theory. The addition of human capital to the neo-classical Solow model, measured by the population's average number of completed school years, was able to account for a large portion of the previously unexplained differences in wealth between nations. Despite the demonstrated importance of education in economic growth and development, little is known about the specific mechanisms that drive growth in the area of educational attainment. What exactly are the forces driving a population's ability and desire to become educated? Or perhaps more importantly, why do some nations exhibit such low levels of educational attainment while others flourish? Does this problem stem from insufficient supply or demand in the market for education? Answers to these types of questions are difficult to reach due to challenges with reverse causality in these relationships, and the difficulty with measuring qualitative factors, such as preferences for education, education quality, and vast number of factors that must be accounted for in analyzing this topic.

There are considerable benefits to advancing knowledge in this topic. More and more, policymakers are focusing their efforts on improving the supply of educational services. In the interest of developing cost-effective and targeted policies to improve educational attainment in developing nations, further research on this topic is essential. Understanding the conditions and factors that influence levels of educational attainment are crucial in making educated decisions on how to address this issue.

Currently, education development initiatives tend to focus on the supply side, commonly centered around improving access and quality of educational services. However, with my analysis I hope to provide a more comprehensive framework involving both supply and demand side factors to model educational attainment. To do this, I provide a case study looking at the developing country of Uganda. The low levels of educational attainment seen in the Ugandan population provide an ideal case to examine deterrents to educational attainment. With the institution of the Universal Primary Education (UPE) program in 1997, the Ugandan government eliminated all primary school fees in an effort to improve education participation and attainment levels (The World Bank, 2007). Despite this change, ten years following the institution of the

UPE program, educational attainment levels are still remarkably low. In 2006, approximately 63% of Ugandans over the age of 15 reported never completing primary school and 20% reported never having attended any formal school (Uganda Bureau of Statistics, 2006a, p.17). These persistent trends beg for further analysis into what exactly is deterring Uganda's population from attaining higher levels of education.

One factor that may be contributing to this issue in Uganda is distance to school. The UNHS 2005/06 Report on the Socioeconomic Module stated that 21.7% of Ugandan students reported living over 3 km away from their primary school, and 5.3% reported living over 5 km away (Uganda Bureau of Statistics, 2006a, p. 23). These distances are made more hindering by low accessibility of motorized transportation, particularly in Uganda's rural areas. This requires the large majority of students to travel these distances on foot (Howe, 2002). This paper intends to assess the impacts of school distance and determine whether they present a significant barrier to educational attainment. I hypothesize that distance to school does in fact coincide with lower levels of educational attainment, due to increased travel costs associated with distance.

My research employs data from the 2005/06 Uganda National Household Survey. Using available data in the survey, I develop a basic framework to model educational attainment. I form three regression models to measure the correlations of various socioeconomic and demographic factors with educational attainment in the Ugandan youth population, as well as in two subsets of this population: current students and non-students. The variables selected for use in my basic model are based on past research models of educational attainment discussed in section II. I supplement this model by introducing a measure of distance to school, school cost, and dummy variables indicating preferences regarding education, and reasons for non-participation. I then compare my results between these samples to determine factors affecting educational attainment in the population as a whole, as well as the relative magnitude of these effects between groups.

I find that in the general population, individual education levels are most significantly correlated with parents' education levels, poverty conditions, urban/rural profile, and pregnancy. I find that the magnitude and direction of distance effects are inconclusive given limitations in the survey data, but formulate a hypothesis for further research. Lastly, looking at the non-student population subset, I find that of all the reported reasons for not attending school included in the model, distance and poor school quality lower expected attainment levels by the greatest margin.

II. DETERMINANTS OF EDUCATIONAL ATTAINMENT

Early models of educational attainment built off of traditional functionalist theory, that is the view that family income and size were the primary determinants of educational outcomes as they provided the financial means necessary to obtain an education (De Graaf, 1986). However, particularly in the developed world, modern education systems providing free universal education have greatly reduced the importance of financial status in determining educational outcomes. While the impact of factors such as sex, race, or wealth, have changed drastically over time, the effects of parental occupation classes and education levels have proven to be a remarkably stable and crucial factor in predicting educational outcomes (De Graaf, 1986). Parent's education level, occupational status, and family presence have all shown to be reliable predictors of their children's outcomes over time (Sander, 1991). In other words, the apple doesn't fall far from the tree. In a US study done using the National Survey of Families and Households, Lehrer (1999) found that education levels of the child's mother and father, as well as their presence in the household, had positive and statistically significant effects on children's educational attainment. A family history of reliance on welfare was also found to have negative and statistically significant effects on education levels (Lehrer, 1999).

Other work has been done drawing the link between religion, culture and educational attainment. Religion has been shown to impact education levels; In the U.S., Jewish individuals exhibited higher attainment levels at all stages of life relative to Protestants and Catholics (Lehrer, 1999) In an attempt to estimate religious and ethnic effects on educational attainment, Sander (1991) determined that differences in economic background and location partly explain correlations between ethnic and religious background and educational attainment, but a residual difference in education levels between ethnic groups remains unexplained. The effects of culture are evidently a challenge when it comes to empirical measurement.

Certain school characteristics and funding have also shown some evidence of impacting educational attainment, with larger expenditures on schooling corresponding with higher levels of education, as well as larger expected earnings for students (Wilson, 2001). Distance has also been found to play a role in school participation, particularly in the developing world (Jensen & Neilsen, 1997). Using Zambia's household survey data, Jensen and Neilson (1997) found that distance to primary schools had a significantly negative effect on school attendance. Similar

effects from distance have been seen in the U.S. and Canada in university populations (Card, 1995) (Frenette, 2006).

The gap in knowledge on this topic is evident, with no widely accepted model or framework in place to even begin mapping the factors at play in this relationship. Based off established work in this field, I construct a model of educational attainment using a top-down approach, beginning with a 29 variable regression model and using a stepwise approach to form a reduced model. This paper contributes a case study of educational attainment in Uganda, a region not yet covered by current research on this topic. I also introduce distance to school as a factor in my model, and assess school distance as a potential barrier to educational attainment in this region. With my analysis, I aim to expand the body of empirical research on this topic to help governing bodies develop effective policy addressing this issue.

III. DATA

My analysis uses data from the 2005-2006 administration of the Uganda National Household Survey, otherwise referred to in this paper as the UNHS 2005/06. The UNHS 2005/06 collected data on approximately 43,000 individuals in 7,400 households, covering all districts of Uganda. The survey's stated objective is to gather reliable data on demographic, social and economic characteristics of the household population to aid in forming national and international development frameworks (Uganda Bureau of Statistics, 2006a). My analysis relies solely on the survey responses reported in the Socioeconomic Module, which gathered data on respondent demographics, education history, socioeconomic indicators, and consumption patterns at the individual and household level.

My analysis focuses on data from the education section in the socio-economic module of the survey (Section 4). In this section, respondents were asked about their education history and background. Individuals that reported being current students were asked for their current year of study, and about various attributes of their school such as yearly expenses, the type of school (private, government, religious etc.), and the distance to school in kilometers. Individuals that reported attending some school in the past, or having never attending school were asked to select a reason from a list of choices for why they didn't attend school, or why they left school. Given the differences in survey methodology and the data provided by the current student and non-

student populations[1], I chose to separate these two populations when analyzing attributes specific to each group. Given this separation based on student status, it is important to note the inherent biases in each group. While this is limiting, this distinction was essential in order to assess factors for which data is only available for a single group. For both groups, I limited my sample to only include data on individuals between the ages of 5-25. I have exclusively chosen this age demographic for two reasons: first, to capture the conditions affecting individuals during their prime schooling years, and second, to reduce estimation error and inaccuracies involved with changes in the Uganda school system over time.

The current Uganda public school system is composed of 7 years of primary school (P.1-P.7) and 6 years of secondary school (S.1-S.6) (Uganda Bureau of Statistics, 2006a). To measure educational attainment, I use the number of school years successfully completed by each individual. The school years completed value is taken from two survey questions asking for "highest grade level completed" (Section 4.4) for non-students[2], and "current grade level" (Section 4.6) for current students[3] (2006c, p. 4). Under this valuation, failed grades are not counted towards the number of years completed.

Some limitations of the dataset include a large number of missing responses among survey questions. This severely limited the sample size available for use in my analysis down to

---

[1] The current student population is composed of all individuals who reported currently attending school at the time the survey was administered. The non-student group includes all individuals who reported attending school in the past and all individuals who reported never attending school at the time the survey was administered.

[2] For non-students, the value for years completed was calculated using a conversion of grade level to years, starting with P.1 as 1 and ending with S.6 as 13. Individuals with less than p.1 or no formal education reported a value of

[2] For non-students, the value for years completed was calculated using a conversion of grade level to years, starting with P.1 as 1 and ending with S.6 as 13. Individuals with less than p.1 or no formal education reported a value of zero. the response "post-primary specialized training or diploma" was inputted as 8 years completed, indicating the completion of 7 years of primary school, and a minimum of one additional year in a post-primary program. The response "post secondary specialized training or diploma" was inputted as 14 years, indicating the completion of 13 years of public school and a minimum of one year in a post-secondary program. The response "completed degree and above" was inputted as 17 years, indicating the completion of 13 years of public school, followed by 4 years of a post-secondary degree program. These estimations are based on the minimum average time to complete such programs, however the accuracy of these values are limited without the ability to accurately account for the full number of years completed for individuals who completed additional education. Fortunately, individuals who reported attending or completing a post-primary or post-secondary specialized or degree program are only captured in a small number of observations within the non-student group.

[3] For current students, the value for years completed was calculated using a conversion of current grade level minus one, converted to years. Note that this value reflects the highest grade level completed, and not the current grade level of the individual. For example, and individual in p.2 would receive a value of 1, p.3 a value of 2, and so on. Current students in post-primary, post-secondary diploma or degree programs were excluded from the sample due to non-comparability of reported distances for these types of education.

2,059 current students and 1,805 non-students. With this, there is also the potential for an unrepresentative sample given the relatively low number of completed surveys. Additionally, it is important to note the difficulties associated with making inferences from self-reported data due to measurement and estimation error. Accordingly, a select number of outlying observations were dropped from the sample due to suspected inaccuracy[4]. Due to the imprecision inherent in self-reported data, careful attention was drawn to ensuring only variables exhibiting highly robust and statistically significant effects were considered in my interpretation of results.

IV. METHODOLOGY

There are three regression models tested in this analysis. First, a basic regression model is formed including all variable measures available for both the current student and nonstudent samples. This basic equation is estimated for the full sample, as well as both the current student and non-student subsets separately. This model is then supplemented with the addition of subsample-specific variables to form two separate augmented models for the current student and nonstudent sample subsets. Each augmented model is regressed on its respective sample subset, and the results are compared to the basic model regression results.

The reason for the use of three separate models is due to the differences in survey methodology between the two sample subsets. Respondents who indicated being current students were asked about their school's characteristics, cost per year, as well as their distance to school. Respondents who indicated that they were not current students were instead asked to report a reason for why they did not attend school or why they left school. Due to the differences in data available for each subset, two separate models were formed specific to each sample subset to supplement the basic model.

BASIC REGRESSION MODEL

The basic model intends to form a simplified predictive model of educational attainment. The selection of variables took a top-down approach; First, I formed a comprehensive list of all potential regressors based on pre-existing theories of educational attainment, then worked through the survey data to find appropriate measures for each variable. In the initial formation of

[4] Observations reporting time spent on household duties larger than 167 hours/week and observations reporting distance to school of greater than 80 km were dropped.

the model, 29 variables were included to account for all factors with a potential to influence educational attainment, given the constraints of the dataset. For the sake of concision, I will only present the variables selected for use in later iterations of the model. Variables that were removed from later models were found either to be consistently insignificant by a large margin, contained unreliable or insufficient data, or were removed due to multicollinearity. A list of all initial variables and the reasons for their removal is available in Appendix C.

The basic model takes the form seen in equation (1) below. Table I shows a full list of the variables included in the basic regression model along with summary statistics.

$$\begin{aligned} Years\ Completed_i = {} & \alpha_i + \beta_1 Father's\ Education_i + \beta_2 Mother's\ Education_i + \beta_3 Hours\ Worked_i + \\ & \beta_4 Time\ Spent\ on\ Household\ Duties_i + \beta_5 Poverty_i + \beta_6 Pregnant + \\ & \gamma Other\ Controls + \varepsilon_i \end{aligned} \tag{1}$$

The independent variables of interest in this regression include the education levels of the mother and father, the number of hours per week reported working in any employment, time spent on household duties, poverty and pregnancy.

The parent's education levels are reported directly in the survey using code values of 1 through 6 (see Table I notes). Consistent with the findings of Sanders (1991) and De Graaf (1986), I would expect the coefficients on both the mother's and father's education level to be positive, reflecting that higher education among parents corresponds with higher education in children.

Hours worked is taken from the number of hours respondents reported working in any employment in the last 7 days (Section 7a). The predicted direction of the relationship between hours worked in employment and school years completed is ambiguous. Higher education levels may correspond with more working hours due to increased job opportunities and skills, however it is also possible that those with little education may work longer hours in a low-income job in order to make ends meet.

TABLE I – BASIC MODEL VARIABLES AND SUMMARY STATISTICS

	Total	Current Students	Non-students
	(1)	(2)	(3)
School Years Completed	4.65	3.35	6.13
	(3.05)	(2.74)	(2.69)
Education level of Father*	2.50	2.58	2.41
	(1.13)	(1.09)	(1.16)
Education level of Mother*	1.94	2.04	1.82
	(0.94)	(0.93)	(0.94)
Hours Worked *(last 7 days)*	13.19	3.84	23.84
	(19.94)	(8.39)	(23.60)
Time spent on household duties *(hours/week)*	23.21	10.48	37.70
	(29.41)	(12.62)	(35.70)
Poverty	1.25	1.26	1.24
	(0.97)	(0.97)	(0.97)
% Pregnant (in last 5 years)	25.78%	0.19%	54.90%
Controls			
Age	16.15	11.74	21.17
	(5.89)	(3.90)	(3.09)
% Primary	77.02%	88.12%	56.57%
% Male	41.67%	48.83%	33.52%
% Urban	23.74%	22.05%	25.65%
Number of observations	3,859	2,054	1,805

Source: Uganda Bureau of Statistics. (2006b). *Uganda National Household Survey 2005/2006.*
Column (1) displays values for the combined population. Columns (2) and (3) display values for only current student and non-student samples, respectively. All columns show means. Standard deviations in parentheses.
Notes: *Both mother's and father's education levels are taken from the following survey code: 1= No formal education, 2= Less than primary, 3= Completed primary, 4= Completed O-level, 5= Completed A-level, 6= Completed university. O-level consists of the first 4 years of secondary school; A-level consists of the last 2 years of secondary school. %Primary shows the percent of respondents with school years completed < 7.

I include a measure of time spent on household duties to reflect potential opportunity costs of school attendance. This value is measured by the aggregate number of hours per week reported fetching water and firewood, cooking for the household, and caring for children and the elderly in the household (Section 7b.6-9). I expect this variable to have a negative relationship with education level, as greater household responsibility increases the opportunity cost of school attendance. Additionally, these household duties are typical of a caregiver role in the family, which does not require higher education levels.

The poverty indicator is derived from the yes/no responses for three survey questions asking if every member in the household has at least two sets of clothes, one pair of shoes, and if every child in the household has a blanket (Section 15.4-6) A response of "no" for each of the above questions adds one point to the poverty indicator. Under this valuation, a "no" response for each question returns a poverty indicator value of 3, and a "yes" response for each question returns a value of zero. This method for indicating the presence of poverty conditions is fairly unsophisticated, but still provides some indication of extreme poverty when present. I expect a negative coefficient on this variable due to the strong, positive relationship between income and education already established in economic literature (Blanden, 2004).

A dummy variable is included for whether the individual reported becoming pregnant in the last 5 years. This intends to capture the negative effects of pregnancy and childbirth on school completion, particularly for young women (Rumbaut, 2007).

A number of control variables are included to account for differences in age, sex, urban/rural residence, and whether the individual's highest education was in primary or secondary school to account for differences in school type. Some variables that would have been valuable additions to the model, but are unavailable in the data include measures of income, community safety and security, racial, cultural and religious factors, and a measure of transportation accessibility.

CURRENT STUDENT AUGMENTED MODEL

The augmented model for current students consists of the basic model variables plus three additional variables: primary school distance, secondary school distance, and school cost. Data on these three measures was only collected for respondents reporting current attendance at school, and therefore this model is applied only to the current student population subset.

The augmented current student model, including the primary and secondary school distance measures and the school cost variable is shown below in equation (2).

$$(2) \quad Years\ Completed_i = \alpha_i + \delta_1 Primary\ School\ Distance_i + \delta_2 Secondary\ School\ Distance_i + \delta_3 School\ Cost_i + \gamma Basic\ Model\ Variables\ and\ Controls_i + \varepsilon_i$$

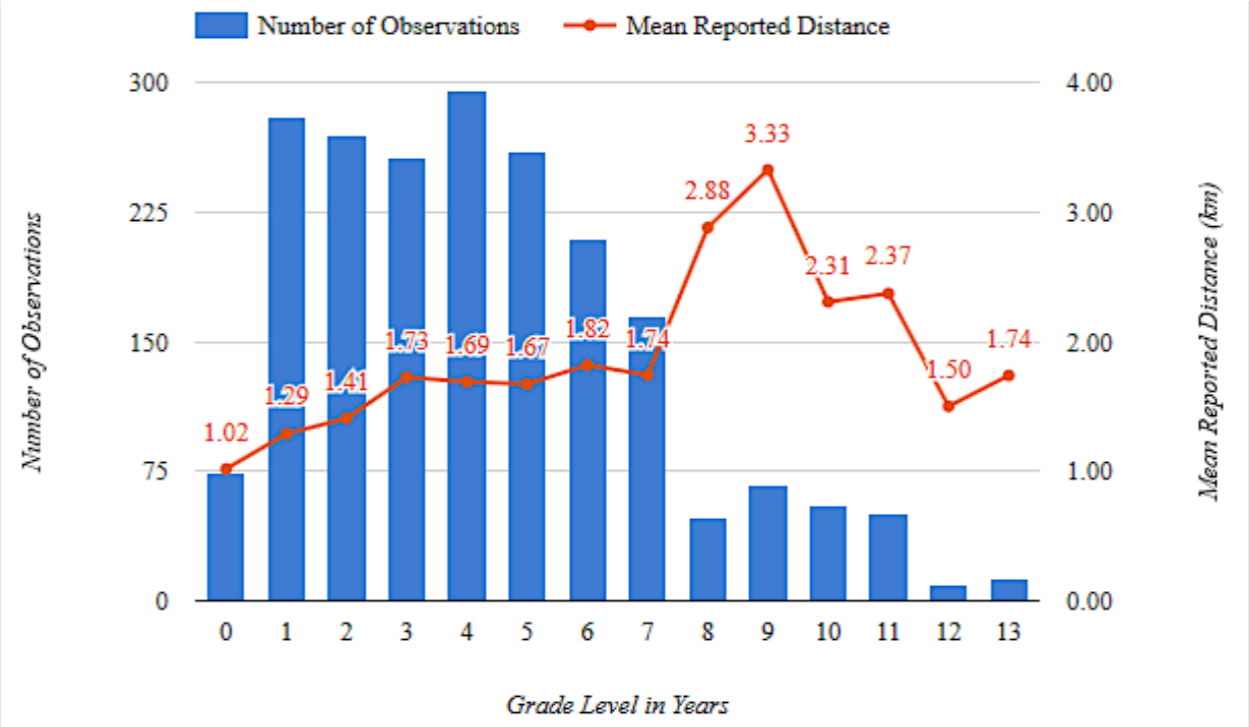
The measures of school distance are taken from self-reported values of distance to school in kilometers (Section 4.9). The primary and secondary school distances are a calculation of the distance values multiplied by a dummy variable for if the individual is in primary or secondary school[5].

I predict that farther distances would impose greater travel costs and increase the burden associated with regular school attendance, as previously suggested in the works of Card (1995) and Frenette (2006). Lengthy commutes reduce the amount of free time an individual has to spend studying, working, or engaging in leisure activities. Due to the costs imposed by distance, I would hypothesize that increased distance to school negatively impacts education participation rates. Moreover, I would expect low mobility in rural areas, where public transportation services are scarce, to further inflate this effect.

While I predict that a negative relationship exists between these variables, this may not be reflected in the estimated distance coefficients, given that we are only looking at a sample of current students. It's possible that the individuals in this group are in school because they are not affected by distance. If so, the estimate may be biased towards zero or statistically insignificant. Moreover, the data shows that average distance to school varies depending on grade level, as seen below in Figure I. Mean reported distance is reported at the grade level, in red. The number of observations in each grade level is shown in blue.

[5] Dummy is coded such that a current grade between 0-7 returns primary=1, otherwise 0. A current grade between 8-13 returns secondary=1, otherwise 0.

FIGURE I – Mean Reported Distance by Grade Level



Source: Uganda Bureau of Statistics. (2006b). *Uganda National Household Survey 2005/2006.* N = 2049

Figure I shows that students in secondary school report larger mean distances to school, with greater variance, than those in primary school. Looking at the number of observations in each year, we see a sharp drop off in respondents at year 8, which is the first year of secondary school. Without controlling for school type, the coefficient on the distance regressor would likely return as a positive value, reflecting that the larger distances seen in secondary school are associated with more years completed. I chose to include primary and secondary school distances as slope dummies to partially control for the differences seen in reported distance based on schooling level.

This model also includes a yearly school cost variable, which intends to measure the effect of financial burdens associated with regular schooling. This variable could show either a positive or negative effect on years completed; negative because of poverty effects, as more expensive schooling becomes less affordable for low-income households. However, this variable could also exhibit a positive coefficient due to the sample consisting of students in both private

schools and government schools. It seems likely that children in private schools, who pay more for their schooling, generally come from wealthier families. These children may also achieve higher levels of schooling as a result of their financial and socio-economic status, forming a positive correlation between school cost and years completed.

NON-STUDENT AUGMENTED MODEL

The non-student group consists of 1,805 individuals between the ages of 5 and 25 who indicated either completing some schooling in the past or none at all. While non-students were not asked to report any measures of school distance, they instead were asked to report a reason for why they chose to leave school if they attended some in the past, or why they chose not to go to school if they never attended. The answer choices that were provided to both groups of non-student respondents (those that completed some school and those that didn't complete any) are summarized in Table II.

TABLE II

REPORTED REASONS FOR LEAVING OR NOT ATTENDING SCHOOL

	Number of Observations	Percent of Population
	(1)	(2)
Too far	10	0.55%
Too expensive	1,005	55.68%
Not willing to attend	124	6.87%
Parents did not want/not allowed	85	4.71%
Had to help at home/with family farm/with family business	56	3.10%
Poor school quality	20	1.11%
Other	505	27.98%
Total	1,805	100%

Adapted from "Code for 3" and "Code for 5" in Section 4 of Uganda National Household Survey 2005/2006: Socioeconomic Questionnaire (p. 4). Column (1) displays the number of observations for each reason. Column (2) shows the percent of individuals in the population who reported the reason. Only answer choices common to both H4Q3 and H4Q5 codes are shown above. The full coded lists of reasons provided to survey respondents are available in Appendix A.

Table II shows that the majority of respondents reported that cost was the determining factor in their decision, followed by preference based reasons, such as being unwilling to attend or their parents not wanting them to attend. A small number of respondents indicated distance or household responsibilities as the leading factor. "Other"[6] reasons not captured by these answer choices also accounted for a large portion of the sample.

To capture the effects of each of these six "reported reasons" as deterrents to education, they are each included as dummy variables in the non-student augmented model regression. A dummy value of one for an answer choice indicates the presence of that choice being selected by a respondent; a value of zero indicates the choice was not selected. Respondents were only allowed to select one choice. This is limiting as it forces individuals to only select the reason that was most influential in their decision, however does not reflect the presence of any other reasons that may have played into the decision. For future research, a survey method which allows for greater flexibility in response to such questions would be helpful for determining the relative size of less prominent deterrents.

With the addition of the reported reason dummies, the non-student augmented model takes the form of equation (3):

$$(3)\quad Years\ Completed_i = \alpha_i + \gamma Basic\ Model\ Variables\ and\ Controls_i + \theta Reported\ Reasons_i + \varepsilon_i$$

V. RESULTS

I estimate the basic model equation (1) using an OLS regression on the full survey sample, as well as both current student and non-student subsamples. The augmented model equations (2) and (3) are estimated using OLS regressions on the current student and non-student subsamples, respectively. Results for the five OLS regressions performed are reported in Table III. Columns 1-3 show estimated coefficients in the basic model regression on the full sample and both subsample groups. Columns (4) and (5) show estimated coefficients in the augmented model regressions for current student and non-student subsamples.

[6] Refers to other options provided in the survey question code not shown in this table. These options were omitted due to differences in the survey methodology. Individuals who never attended any school were given a different list of options from individuals who attended some school in the past. As a result, only options provided to all respondents were included in the regression model to ensure the correct coding of the dummy variables. The full coded lists of reasons provided to survey respondents is available in Appendix A.

TABLE III
Results of OLS Regressions on Years of School Completed

	Basic Model			Augmented Model	
	Full Sample	Current students	Non-students	Current students	Non-students
	(1)	(2)	(3)	(4)	(5)
Education level of father	0.191***	0.162***	0.204***	0.159***	0.209***
	(0.026)	(0.028)	(0.038)	(0.028)	(0.038)
Education level of mother	0.153***	0.089***	0.232***	0.070**	0.231***
	(0.031)	(0.034)	(0.048)	(0.034)	(0.048)
Hours worked	-0.003**	-0.005	0.003	-0.004	0.002
	(0.001)	(0.003)	(0.002)	(0.003)	(0.002)
Time spent on household	0.000	0.003	-0.004***	0.004*	-0.004***
duties	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Urban	0.194***	0.151**	0.147*	0.095	0.119
	(0.061)	(0.070)	(0.090)	(0.071)	(0.090)
Poverty	-0.172***	-0.139***	-0.193***	-0.116***	-0.183***
	(0.028)	(0.030)	(0.042)	(0.030)	(0.042)
Pregnant	-0.864***	-1.639***	-0.405***	-1.644***	-0.408***
	(0.092)	(0.622)	(0.135)	(0.619)	(0.134)
Male	-0.227***	-0.016	-0.511***	-0.007	-0.226***
	(0.061)	(0.056)	(0.136)	(0.056)	(0.136)
Primary school distance				0.016	
				(0.015)	
Secondary school distance				-0.016	
				(0.020)	
School cost				0.00***	
				(0.00)	
Reported Reason Dummies					
Too expensive					-0.226***
					(0.087)
Too far					-1.556***
					(0.511)
Not willing					-0.451***
					(0.163)
Parents did not want					-0.609***
					(0.189)
Had to help at home					-0.481**
					(0.227)
Poor school quality					-1.282***
					(0.365)
Number of observations	3,859	2,054	1,805	2,054	1,805
Adjusted R^2	0.745	0.799	0.643	0.801	0.648

Columns 1-3 show estimated coefficients in the basic model regression on the full sample and both subsample groups. Columns (4) and (5) show estimated coefficients in the augmented model regressions for current student and non-student subsamples. Standard errors in parentheses are listed below: *significant at 10%, **significant at 5%, ***significant at 1%. *Note*: Results for control variables *Age* and *Primary* not included in table. See Appendix B for regression coefficients and standard errors.

The education levels of both mother and father were both found to be positively correlated with years completed for all groups, significant at the one percent level. The poverty and pregnancy indicators were also found to have significantly negative coefficients in all samples. Hours worked showed a small, but significant negative correlation with school years completed in the basic regression on the full sample, but was insignificant in all other cases. In the basic model, time spent on household duties was only found to be significant for the non- student population, showing a negative coefficient. The same effect was seen for non-students in the augmented model regression, but a positive coefficient was seen for current students. Residing in an urban region showed a positive and significant correlation with years completed in all three samples of the basic regression, however this became insignificant in the augmented model regressions. The male dummy showed a highly significant, negative coefficient in the non-student subsample and full sample, but this relationship was insignificant for current students in both basic and augmented model regressions.

The coefficients on both primary school distance and secondary school distance were found to be insignificant. School cost reported a highly significant coefficient of zero, indicating that school cost had no correlation with years completed. The addition of distance and school cost to the model increased the significance of the time spent on household duties coefficient to the 10% level, and made the urban dummy insignificant. The augmented model also marginally increased the adjusted R squared value to 0.801.

In the augmented model regression on the non-student population, all reported reason dummy variables were found to be highly significant and negatively correlated with school years completed. Of all the reported reason dummies, “too far” was found to have the largest negative coefficient, followed by “poor school quality”. The weakest negative coefficient was found for the “too expensive” reason dummy. The addition of these dummies to the basic model marginally increased the adjusted R squared value to 0.648.

VI. DISCUSSION

Perhaps the most remarkable of these findings is the robustness of the estimated coefficients on parent’s education, pregnancy, and poverty. The positive and highly significant coefficients on parent’s education conform to my hypothesis that a child’s level of educational

attainment is positively correlated to the attainment levels of its parents. This effect was seen across the board in all samples and regression models.

The negative and highly significant coefficient on the poverty indicator is consistent across all groups. Consistent with my predictions, poverty is negatively correlated with number of school years completed, with this relationship being strongest for non-students.. To supplement this result, the coefficient on the reported reason of "too expensive" also holds a highly significant negative relationship with school years completed. This study does not establish the direction of causality in this relationship, but these findings confirm that poverty and financial struggle are still closely tied to educational attainment in Uganda.

The weak and somewhat inconclusive effects found for time spent on household duties are not entirely surprising. The values reported on hours spent per week on these types of activities are difficult to estimate and unlikely to be accurate. Without reliable data, there is little that can be concluded about the relationship between time spent in these alternative activities and school completion.

Although I did not set out to measure gender differences in educational outcomes, I found that gender was only significant for samples, which included non-students. For individuals in school, there was no significant effect between gender and years completed. However, in the full sample and non-student subsample, a negative correlation was seen between being a man and school years completed. This seems contrary to what you would expect. It's possible that this is due to a slight bias in the non-student subsample, which is only 33.52% male, or potential multicollinearity with the pregnancy dummy.

The zero coefficient on school cost in the current student regression suggests that the observed poverty effects may go beyond affordability of school. If the price of education was the limiting factor for individuals in poverty, we would expect school cost to be negatively related to years completed. This could be simply a result of bias in the current student group, indicating that those that are most affected by cost have been excluded from the sample, leaving only those that are unaffected by cost. Another possibility is that a positive correlation between cost and years completed from more affluent private school students is counteracting the negative relationship from poverty seen in the rest of the sample. It is impossible to know for sure given the limitations in data available in this study, but the sheer number of non-students who reported

cost as the prevailing reason for their decision to not attend school suggests there may be an omitted variable or selection bias in this sample.

Selection bias between groups also makes the interpretation of the distance coefficients quite difficult. Distance was found to be insignificant in all cases, even after separating primary and secondary school distances. However, I am not convinced that much can be gleaned from this result given the selection bias in this sample, and the lack of comprehensive data in this study on school distances throughout a child's entire education history.

I suspect that distance would play the largest role in education decisions made at the points of transition from no school to primary school, and from primary school to secondary school. At the initial decision node, individuals with no education choose whether or not to attend primary school. At this point, a self-selection occurs where individuals that are sufficiently deterred by distance choose not to enter the student population. Those that enter primary school face the same barrier of distance on a daily basis, but their decision to enter primary school indicates that the effect of distance was sufficiently small enough to still receive positive net benefits from attendance. Similarly, at the point of transition from primary to secondary school, another self-selection would occur. Those that are sufficiently deterred by the larger average distances to secondary schools would self-select out of the secondary student population, leaving only individuals that are less influenced by the effects of distance. This phenomenon may partially explain the large dropout rates seen upon entry to secondary school in Figure I. However, without comprehensive data on distances affecting all individuals, both students and non-students, we cannot know for sure if this is the case.

The effects of the "too far" reported reason do provide a small glimpse into distance's effects on non-students. The coefficient estimate of -1.556 on the "too far" dummy showed the strongest negative relation to school years completed of all the reported reason dummies. This finding supports my hypothesis that individuals who are substantially inhibited by distance are more likely to be seen within the non-student population due to a self-selection out of the student population.

VII. CONCLUDING REMARKS

More generally, from these findings we can conclude that there are evidently major differences between the composition of the current student and nonstudent populations. While

we may not be able to fully measure these differences, it is clear that there is a complex and dynamic system of factors determining educational attainment. This paper provides empirical evidence confirming the role of established socioeconomic factors, and introduces a new hypothesis about the effects of school distance in determining educational outcomes. My findings confirm the importance of parents' education levels, pregnancy, and poverty in educational outcomes. Ambiguity surrounding the effects of school distance, cost, and other school characteristics seen in my results call for a deeper analysis in later research.

As discussed throughout the paper, there are limitations in what can be concluded from this study. The inconsistent survey methodology between student and nonstudent groups in this dataset pose the largest limitation in my interpretation of results. An analysis of a dataset with in- depth accounts of educational histories in an unbiased population, as well as comprehensive data on school characteristics is needed before any concrete claims can be made about distance as a barrier to education. A replication of this model in a different population dataset would be an important next step in order to compare whether the effects seen here can be generalized to a more broad-based model of educational attainment.

REFERENCES

Blanden, J. (2004). Family Income and Educational Attainment: A Review of Approaches and Evidence for Britain [Abstract]. *Oxford Review of Economic Policy, 20*(2), 245-263. doi:10.1093/oxrep/grh014

Card, D. 1995. "Using geographic variation in college proximity to estimate the return to schooling". In Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, Toronto: University of Toronto. doi:10.3386/w4483

Cohen, A. (1991). Dummy Variables in Stepwise Regression. *The American Statistician*, *45*(3), 226–228. http://doi.org/10.2307/2684296

De Graaf, P. M. (1986). The Impact of Financial and Cultural Resources on Educational Attainment in the Netherlands. *Sociology of Education, 59*(4), 237-246. doi:10.2307/2112350

Frenette, M. (2006). Too Far To Go On? Distance to School and University Participation. *Education Economics, 14*(1), 31-58. Retrieved February 15, 2016, from http://www.tandfonline.com

Howe, J. (2002). Boda Boda - Uganda's rural and urban low-capacity transport services. Final report. Retrieved February 21, 2016, from http://r4d.dfid.gov.uk/Output/5453/

Jensen, P., & Nielsen, H. S. (1997). Child labour or school attendance? Evidence from Zambia. *Journal of Population Economics, 10*(4), 407-424. doi:10.1007/s001480050051

Lehrer, E. L. (1999). Religion as a Determinant of Educational Attainment: An Economic Perspective. *Social Science Research, 28*(4), 358-379. doi:10.1006/ssre.1998.0642

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A Contribution to the Empirics of Economic Growth. The Quarterly Journal of Economics, 107(2), 407-437.

Rumbaut, R. G. (2005). Turning points in the transition to adulthood: Determinants of educational attainment, incarceration, and early childbearing among children of immigrants. *Ethnic and Racial Studies, 28*(6), 1041-1086. doi:10.1080/01419870500224349

Sander, W. (1992). The effects of ethnicity and religion on educational attainment. *Economics of Education Review, 11*(2), 119-135. doi:10.1016/0272-7757(92)90003-l

The World Bank. (2007). A Better Managed Education System in Uganda. Retrieved February 24, 2016, from http://go.worldbank.org/DDP4HYKHJ0

Uganda Bureau of Statistics. (2006a). *Uganda National Household Survey, 2005/2006: Report on the Socio-economic Module*. Retrieved from http://www.ubos.org/unda/index.php/catalog/ 12/study-description#page=accesspolicy&tab=study-desc

Uganda Bureau of Statistics. (2006b). *Uganda National Household Survey 2005/2006* [Data file]. Retrieved from https://openmicrodata.wordpress.com/2010/02/12/uganda-national-household-surveys-ihs-unhs

Uganda Bureau of Statistics. (2006c). *Uganda National Household Survey 2005/2006*: *Socioeconomic Questionnaire.* Retrieved from https://openmicrodata.wordpress.com/2010/02/12/uganda-national-household-surveys-ihs-unhs

Uganda, Planning and Economic Development, Ministry of Finance. (2013). *Millennium development goals report for Uganda 2013: Special theme: Drivers of MDG progress in Uganda and implications for the post-2015 development agenda* (pp. 1-94). Kampala. Retrieved April 15, 2016, from

http://www.undp.org/content/dam/undp/library/MDG/english/MDG Country Reports/Uganda/Uganda MDG Report-Oct 2013.pdf

Wilson, K. (2001). The Determinants of Educational Attainment: Modeling and Estimating the Human Capital Model and Education Production Functions. *Southern Economic Journal, 67*(3), 518. doi:10.2307/1061450

APPENDIX TABLES

APPENDIX A
Codes for S4.3 AND S4.5

S4.3: Why have you not attended school? (1)	S4.5: Why did you leave school? (2)
1= Too expensive	1= Completed desired schooling
2= Too far away	2= Further schooling not available
3= Poor school quality	3= Too expensive
4= Had to help at home	4= Too far away
5= Had to help with farm work	5= Had to help at home
6= Had to help with family business	6= Had to help with farm work
7= Education not useful	7= Had to help with family business
8= Parents did not want	8= Poor school quality
9= Not willing to attend	9= Parents did not want
10= Too young	10= Not willing to attend further
11= Orphaned	11= Poor academic progress*
12= Displaced	12= Sickness or calamity in family
13= Disabled	13= Pregnancy
14= Insecurity	14= Other (specify)
15= Other (specify)	

Adapted from "Code for 3" and "Code for 5" in Section 4 of Uganda National Household Survey 2005/2006: Socioeconomic Questionnaire (p. 4). Column (1) displays code for Section 4.3 and column (2) displays code for Section 4.5.

APPENDIX B

Full Variable List, Survey Reference, and Derivation

Variables (1)	Survey Question (2)	Value Derivation (3)
Years of school completed	S4.4: What was the highest grade you completed? S4.6: What grade are you currently attending?	See footnote 2
Education level of father	S3.4: What was/is the highest level of father's education completed?	1= No formal education 2= Less than primary 3= Completed primary 4= Completed O-level 5= Completed A-level 6= Completed university
Education level of mother	S3.2: What was/is the highest level of mother's education completed?	1= No formal education 2= Less than primary 3= Completed primary 4= Completed O-level 5= Completed A-level 6= Completed university
Hours worked	S7a.10-16: During the last 7 days, how many hours did you work on each day?	Sum of responses
Time spent on household duties	S7b.6: During the past 7 days, how many hours did you spend fetching water for the household including travel time? S7b.7: During the past 7 days, how many hours did you spend fetching firewood for the household including travel time? S7b.8: During the past 7 days, how many hours did you spend in cooking for the household? S7b.9: During the past 7 days, how many hours did you spend in taking care of children and the elderly?	Sum of responses
Poverty	S15.4: Does every member of the household have at least two sets of clothes? S15.5: Does every child in this household (all those under 18 years old) have a blanket? S15.6: Does every member of the household have at least one pair of shoes?	Yes = 0, No = 1 Sum values to yield poverty indicator value between 0 and 3.
Pregnancy dummy	S6.12: Has [NAME] ever been pregnant during the last 5 years?	Yes = 1, No = 0
Age	S2.9: How old is [NAME] in completed years?	Reported directly
Male dummy	S2.4: Sex	Male = 1, Female = 0
Urban dummy	Indicated by geographic region in Section 1	1 = Urban 0 = Rural
Primary/Secondary dummies	S4.4: What was the highest grade you completed? S4.6: What grade are you currently attending?	Current grade between 0-7 returr primary=1, otherwise 0. Current grade between 8-13 returns secondary=1, otherwise 0.
Distance to school	S4.9: Distance to the school in km?	Reported directly
Primary/Secondary school distance	S4.9: Distance to the school in km? S4.6: What grade are you currently attending?	Distance value multiplied by primary or secondary dummy
Reported Reason Dummies	S4.3: Why have you not attended school? S4.5: Why did you leave school?	Reason present = 1 Reason absent = 0

Adapted from Uganda National Household Survey 2005/2006: Socioeconomic Questionnaire.

APPENDIX B2

Control Variable Regression Results Omitted from Table IV

	Basic Model			Augmented Model	
	Full Sample	Current students	Non-students	Current students	Non-students
	(1)	(2)	(3)	(4)	(5)
Age	0.254***	0.429***	0.091***	0.426***	0.089***
	(0.006)	(0.009)	(0.013)	0.009	(0.091)
Primary	-3.765***	-3.223***	-3.767***		
	(0.067)	(0.104)	(0.083)		

Columns 1-3 show estimated coefficients in the basic model regression on the full sample and both subsample groups. Columns (4) and (5) show estimated coefficients in the augmented model regressions for current student and non-student subsamples. Standard errors in parentheses are listed below: *significant at 10%, **significant at 5%, ***significant at 1%.

Note: For current students, the Primary dummy indicates whether current grade level is at the primary school level. For non-students the dummy indicates whether the grade following their highest grade completed is at the primary school level.

APPENDIX C – List of Variables Removed from Regression Model

	Reason for Removal
Demographic Factors	
Married dummy	Incomplete/Insufficient data
Household size	Insignificant in all samples
Mother dummy	Insignificant in all samples
Father dummy	Insignificant in all samples
Socioeconomic Factors	
Owns transportation equipment	Incomplete/Insufficient data
Income Tax	Incomplete/Insufficient data
Financial aid dummy	Incomplete/Insufficient data
Health and Safety	
Days off for personal illness	Incomplete/Insufficient data
Days off for family illness	Insignificant in all samples
Civil Strife dummy	Insignificant in all samples
Crime dummy	Insignificant in all samples
Racial/Ethnic tension dummy	Insignificant in all samples
Reported Reason Dummies	
Education not useful	Incomplete/Insufficient data
Disabled	Incomplete/Insufficient data
Sickness or calamity in family	Incomplete/Insufficient data
Pregnancy	Incomplete/Insufficient data
School Characteristics	
Government school dummy	Insignificant in all samples
Private school dummy	Insignificant in all samples
NGO school dummy	Insignificant in all samples