WIDER Working Paper No. 2013/065

Class size versus class composition

What matters for learning in East Africa?

Sam Jones*

June 2013

Abstract

Raising schooling quality in low-income countries is a pressing challenge. Substantial research has considered the impact of cutting class sizes on skills acquisition. Considerably less attention has been given to the extent to which peer effects, which refer to class composition, also may affect outcomes. This study uses new microdata from East Africa, incorporating test score data for over 250,000 children, to compare the likely efficacy of these two types of interventions. Endogeneity bias is addressed via fixed effects and instrumental variables techniques. Although these may not fully mitigate bias from omitted variables, the preferred IV results indicate considerable negative effects due to larger class sizes and larger numbers of overage-for-grade peers. The latter, driven by the highly prevalent practices of grade repetition and academic redshirting, should be considered an important target for policy interventions.

Keywords: East Africa, education, peer effects, class size
JEL classification: J01, I21, I25, I28

University of Copenhagen, email: Sam.Jones@econ.ku.dk

This study has been prepared within the UNU-WIDER project 'ReCom—Research and Communication on Foreign Aid', directed by Tony Addison and Finn Tarp.

UNU-WIDER gratefully acknowledges specific programme contributions from the governments of Denmark (Ministry of Foreign Affairs, Danida) and Sweden (Swedish International Development Cooperation Agency—Sida) for ReCom. UNU-WIDER also gratefully acknowledges core financial support to its work programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

ISSN 1798-7237 ISBN 978-92-9230-642-7

Acknowledgements

I am gratfeul to Miguel Niño-Zarazúa for very helpful comments and encouragement. Thanks also to Youdi Schipper, Sara Ruto and Rakesh Rajani, plus others in the Uwezo team, for access to the data and collaboration. All errors of omission or commission are the author's.

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by the author.

1 Introduction

Concerns regarding the quality of schooling in developing countries are not new. Fuller (1986) notes that while real expenditures per pupil increased significantly in high-income countries during the 1970s, per pupil expenditures declined in low-income countries over the same period. Thus, by 1980, for every US$1 spent on a pupil in low-income countries, pupils in higher income countries received US$31. In part stimulated by the Education for All campaign, launched in 1990, the last two decades have seen a resurgence of attention to primary education in the developing world. This study focuses on the empirical case of East Africa, defined here as comprising Uganda, mainland Tanzania and Kenya, a region where education challenges and changes have been significant. As late as 1999, official figures compiled by UNESCO (2011) show that net enrolment rates in primary education were as low as 49 percent in Tanzania and 63 percent in Kenya. However, bolstered by the abolition of user fees in Uganda in 1997, Tanzania in 2002 and Kenya in 2003, the vast majority of children now attend primary school in these countries. As of 2009, official estimates of net enrolment were 83 percent in Kenya, 97 percent in Tanzania and 92 percent in Uganda (UNESCO 2011).

Expansion of access has been accompanied by growing concerns regarding quality. Nishimura et al. (2008), for example, notes that in Uganda between 1997 and 2004, primary school enrolment increased by 141 percent while the number of teachers and schools increased by 41 percent. Negative impacts such as larger class sizes, inadequate class rooms and poorly educated teachers are frequently attributed to the introduction of universal primary education (UPE). One author quotes a Ugandan brick layer as saying: ' ... children in UPE schools can neither read nor write their names, yet they keep on being promoted to higher classes. UPE promotes failures, for example, a child who scores 80 marks out of 400 can take the 12th position out of 600 pupils. These are all failures and yet they are promoted to the next class.' (Okuni 2003: 34). The point is that standard measures of performance in the sector, including access rates and grade attainment, may give a false impression of the amount of learning that is being achieved.

The need to raise learning outcomes via improvements to school quality is widely acknowledged in East Africa. Uganda's recent National Development Plan identifies 'decongesting' overcrowded classrooms and using class size indicators as a basis for disbursing funds as some of the key interventions to raise primary school quality (Republic of Uganda 2010: 220). While, few would reject the diagnosis that average learning outcomes are unsatisfactory (for further evidence see Uwezo 2012), the deeper question is what policy makers can and should do about it. A natural focus of policy interventions is to augment school resources and, in particular, reduce

class sizes. Reflecting gains in enrolment as well as demographic trends, both casual observers and official statistics agree that classes in government primary schools across the region are large by international standards, typically containing at least 50 pupils (see UNESCO 2011). Conventional wisdom holds that classes of more than 40 pupils per teacher are detrimental to learning. However, rigorous evidence on the effects of larger class sizes is more guarded in its conclusions. In developed countries there is a growing consensus that reducing class sizes is not a cost-effective policy intervention. For developing countries, where high quality evidence remains limited, school resources including smaller classes appear to be more strongly associated with learning outcomes; even so, any positive impacts from smaller classes are often undermined by weak incentive systems facing teachers.

School resources are not the only determinants of learning outcomes. Peer effects are one set of factors that also may be amenable to public policy interventions (Sacerdote 2011). In principle, they are applicable to challenges in East Africa. For example, grade retention and late enrolment are widespread, producing classrooms that contain pupils of very wide ranging abilities and ages. In other contexts, these practices have been criticized for raising the per pupil cost of a completed primary education and reducing final grade attainment. Classes composed of highly diverse types of pupils also have been found to generate particular pedagogical challenges. These concerns are reflected in academic and popular debates about education in East Africa, motivating their inclusion here.

Following the above, the aim of this study is to examine the learning effects associated with: (i) the size of classes in which pupils learn; and (ii) the share of overage pupils among a child's classmates. The intention is to identify whether these effects are significant, whether they are of an economically meaningful magnitude, and whether these effects vary within the region – e.g., between countries or between population sub-groups. In doing so, three main contributions are anticipated. The first is to add to the evidence base regarding the determinants of learning outcomes in sub-Saharan Africa (hereafter SSA). Second, rather than focus on one intervention in isolation, I aim to compare the magnitudes of class size versus peer age effects. As Hattie (2002) opines, the issue that needs to be addressed is not whether one or other factors have a statistically significant impact on learning. The matter of arguably greater weight concerns the relative importance (and thus cost/benefit) of addressing different types of constraints. Third, to the author's knowledge, the magnitude of peer effects on learning have not been quantified in a low-income SSA context using large scale survey micro-data. This study thus adds to the debate regarding the efficacy of grade retention practices.

The remainder of the paper is structured as follows: Section 2 briefly reviews previous research. Section 3 introduces the database and the measure of cognitive ability, which is the primary outcome of interest. Section 4 discusses the analytical methods to be used, giving attention to how endogeneity bias in the class size and peer overage effects might be addressed. Following previous literature, instrumental variables techniques that rely on variation at the district and school levels are proposed. Section 5 presents the results, starting with region-wide estimates and moving down to population sub-groups. The main findings are that fixed effects methods are likely to be unreliable due to remaining bias from measurement error and unobserved grade-specific variables that affect learning. The instrumental variables results are preferred and indicate statistically significant negative effects from larger class size and from having more overage-for-grade peers. Interpretated in light of the multiple children who are expected to benefit from changes in class conditions, the magnitude of both these effects is material. At the same time, the evidence suggests that reducing the prevalance of overage-for-grade pupils may often be more cost-effective than reducing class sizes. Section 6 concludes.

2 Previous research

An extensive literature considers the determinants of education outcomes, both within and between countries. Following the notion of an educational production function (see Section 4), a wide range of factors has been postulated to influence the learning process. These include the innate characteristics of individuals, family background factors (e.g., parental levels of education), school resource inputs (e.g., availability of learning materials), aspects of school or teacher quality (e.g., teachers' years of experience) and environmental factors (e.g., peer or neighbourhood effects). Different studies often focus on the role of specific types of factors, and school resource inputs have been studied with particular frequency. This is understandable. Of the numerous factors that determine learning, quantitative school resources are among the most easily manipulated by a bureaucracy through changes to school budgets and spending criteria. Attention to school resource effects is further reinforced by public attitudes, which point to a widespread conviction that children's learning outcomes can be directly improved by putting more resources into schools, especially via smaller class sizes.

It is impossible to do justice to the vast number of studies that consider the effects of increased schooling inputs on educational outcomes. Nonetheless, following the publication of research from an increasing number of countries, there is growing evidence to support the viewpoint that

quantitative resource improvements have, at most, a weak positive effect on learning. Taking the case of the USA, Hanushek et al. (2003) note that educational expenditures per pupil have risen continuously over time, more than doubling in real terms from 1970 to 2000. Although this increase has been reflected in measures of school resources, such as a fall in the pupil-teacher ratio from 22.3 to 16.0 over the same period, indicators of actual learning outcomes reveal no clear evidence that these additional resources have been effective. Rigorous econometric studies support this conclusion. Cho et al. (2012) use random variation in births in local school catchment areas of Minnesota as an exogenous source of variation in primary school class sizes. This addresses the endogeneity of class size and leads the authors to conclude that the positive impact of smaller classes is tiny. Indeed, based on a review of a wide range of studies and echoing the earlier conclusion of Hattie (2002), Chingos (2012) suggests that 'many school systems in the United States have overinvested in class-size reduction and that increasing class size in some situations may represent a budget-cutting strategy that minimizes harm to students.' (p. 2).

Similar findings emerge from studies undertaken outside the USA. Based on a cross-section of countries participating in the OECD's Programme for International Student Assessment (PISA) tests,[1] Hanushek and Wößmann (2011) find a weak positive unconditional association between expenditure per pupil and test score outcomes. However, this association is rendered statistically insignificant once the authors control for differences in national income levels. Wößmann (2005) considers class size effects using on The Third International Mathematics and Science Study (TIMSS), focussing on students in lower secondary education across 17 West European countries. Using various techniques to address identification problems, he finds no consistent evidence that smaller classes are associated with higher test scores. However, whilst the effects of access to other resources, such as textbooks, appear positive in some instances, no consistent pattern is found across countries. Using an expanded sample of 39 countries, Wößmann (2003) argues that differences in test scores between countries are principally attributable to differences in institutional arrangements (e.g., incentive structures) not resource inputs.

The above studies refer to research for middle- and high-income countries. Evidence from low-income countries is less extensive. Nonetheless, two broad conclusions can be drawn from available studies. First, evidence from retrospective (observational) surveys suggest that quantitative school resource inputs tend to extert a comparatively stronger (positive) effect in lower- as compared to higher-resource settings (for an early statement of this thesis see Heyneman and Loxley 1983). For example, Case and Deaton (1999) consider variation in

[1] See http://www.oecd.org/pisa.

class sizes and other resources driven by racial segregation of schools under apartheid in South Africa. They find no effect of variation in class sizes for white children, but significant negative effects of larger classes for black children and particularly with respect to numeracy outcomes. Based on data from a range of Southern and Eastern African countries, Fehrler et al. (2009) find negative effects of class sizes above a threshold of around 60 pupils, as well as positive effects from textbook availability (see also Michaelowa 2001). These results are not implausible. One explanation is that school inputs display diminishing marginal returns (Figlio 1999) – thus, weaker effects found in developed countries arise because they operate on a much flatter portion of the curve. An alternative (but complementary) explanation is that school inputs substitute for teacher quality (Hanushek and Wößmann 2011). The notion is that higher quality (better motivated) teachers are able to achieve effective learning outcomes independent of the resources available to them. In contrast, poorer quality teachers rely more heavily on available inputs and are less proficient in adapting their pedagogical approach to overcome resource constraints.

Second, evidence from randomized field experiments (prospective studies) has considerably nuanced the findings from studies that rely on retrospective data. As summarized in Kremer and Holla (2009), while the former studies show that school inputs *can be* effective in supporting learning outcomes, significant and positive effects tend to arise only in contexts where teachers face appropriate incentives to ensure that learning takes place. If more resources are simply added on top of badly performing or distorted systems, few improvements are found. Duflo et al. (2012), for example, report an experiment that provided Kenyan primary school governance committees funds to hire a (low-cost) local teacher on a short-term contract, thereby reducing class sizes from around 80 to 40 pupils in treatment schools. Without additional interventions to mitigate negative compensating behaviours by teachers, however, no benefits from smaller classes were found for pupils randomly assigned to incumbent civil service teachers in treatment schools. Students assigned to contract teachers scored 18 percent of a standard deviation higher on tests compared to students assigned to civil service teachers in the same treatment schools.[2] Even so, persistent and stronger results from the programme were only found for students assigned to a contract teacher in schools whose governance committees received training on how to monitor teacher performance.

Tackling observable shortages in school resources constitutes one kind of educational intervention. Less attention has been given to the impacts of organizational changes that alter the

[2]See Bold et al. (2012) for comparable results from a scaled-up version of a similar contract teacher experiment in Kenya; they only find statistically significant results from schools subject to an intervention managed by an NGO rather than the government.

composition or average characteristics of pupils in a given classroom. Class composition can be considered important because of the presence of peer effects, which can be broadly defined as occurring when the presence of student type A in the classroom affects the educational outcome of student B (for an overview see Hattie 2002; Epple and Romano 2011). Peer effects can take direct and indirect forms; they also extend to cases where particular combinations of children in the same classroom render certain teaching methods less effective. To the degree that these effects are material, it follows that either establishing or avoiding such combinations of pupils in the same class (e.g., via tracking) could yield large effects on learning.

Recent summaries of the empirical literature, which also predominantly refer to studies undertaken in the USA or Western Europe (see Sacerdote 2011; Epple and Romano 2011), indicate that peer effects are statistically significant and occur in numerous forms (e.g., through gender composition, racial and class disruption effects). As noted in the introduction, the pertinent aspect of these debates to low-income (SSA) contexts is that primary school classrooms frequently contain pupils of extremely varied initial abilities and ages. This reflects a rapid expansion of access to primary schooling over the past decade, limited access to pre-school services, a historical legacy of unequal access to education, significant numbers of late starters, and varied practices regarding grade retention. The sheer diversity of pupils in classrooms (see further below) suggests that peer effects could be material, implying that actions to address them could offer an alternative means to boost learning outcomes. Rigorous evidence for such effects in low-income contexts, however, is almost non-existent. An exception is Duflo et al. (2011) who report results from the same Kenya Extra Teacher Programme field experiment noted above, but focus on a random subset of treatment schools which tracked children into smaller classes by prior ability. Contrary to concerns that tracking principally benefits high-achieving students due to direct peer effects and may even hurt low-achievers (see Betts and Shkolnik 2000), the authors find test score gains across the distribution of initial ability for children subject to tracking. The proposed mechanism is that, by grouping students by initial ability, teachers are able to focus instruction toward the median level of ability in the class.

Establishing tracking in resource-constrained settings may not be feasible in many instances, particularly for smaller rural schools. Nonetheless, the notion that class composition influences the efficacy (and effort) of teachers has broader implications. A small but significant literature addresses the benefits and costs of grade retention, whereby low-achieving pupils are prevented from progressing to higher grades, typically when they fail to surpass an ability threshold. As Brophy (2006) summarizes, there is mounting evidence that the longer run costs of grade retention exceed any short-term benefits. Focussing on the effects on those children held back, a

number of studies credibly find that repeaters are more likely to drop out and have lower levels of final attainment (e.g., Roderick et al. 2005; Manacorda 2012). Comparable negative direct effects have been confirmed in the context of SSA primary schools, where grade repetition is widespread (Motala 1995; Michaelowa 2003; Ndaruhutse 2008; Glick and Sahn 2009).

In addition to failing to promote the development of human capital for held back pupils, evidence also suggests that grade retention introduces significant (resource) efficiency costs and may indirectly harm non-retained students. Lavy et al. (2012), for example, show that a higher proportion of repeaters in Israeli middle- and high-schools has a negative effect on pedagogical practices and student-teacher interactions, impairing learning for low-achieving regular pupils in particular. These kinds of effects are germane in low-income SSA not only because of the prevalence of grade retention, but also because many children begin school at a late age (Wells 2009). This practice, sometimes described as academic redshirting, is often adopted by parents (especially of boys) in the belief that older children will outperform their younger classmates. However, the available evidence is mixed that children who delay entrance achieve any educational gains from doing so (for negative conclusions see Deming and Dynarski 2008; Black et al. 2011). Moreover, in light of previous discussion, redshirting in the context of substantial repetition may further exacerbate any unhelpful classroom dynamics associated with having an uneven age composition of peers.

3 Data

3.1 Uwezo surveys

In response to growing concerns about the quality of education across East Africa, as well as an absence of public data that might be used to monitor such trends, Uwezo (meaning 'capability' in Kiswahili) was established in 2009 to monitor and boost public awareness about levels of literacy and numeracy in the region. Inspired by a similar exercise carried out in India since 2005 by the Assessment Survey Evaluation Research Centre (ASER), a core activity has been the design and implementation of large-scale household surveys. To date, three rounds of the Uwezo surveys have been completed. These do not constitute panel data and the second round survey from 2011 (Uwezo 2) is exclusively used here. This is because the first round survey had a more limited geographical coverage and the third round data (from 2012) is now being processed for future analysis. The Uwezo 2 data provides representative samples from 75 percent of all

Kenyan districts, 100 percent of all Tanzanian districts and all but one district in Uganda.[3]

The data collection process used by Uwezo deserves mention. Sample design was provided by the official national statistics agency in each country based on population census frames. This ensures the data is representative at the national and district levels for all children of school age (see below) who are resident in households at the time of the survey. Following the approach adopted and refined by ASER, local community groups were entrusted with the enumeration of the survey in each district. Specifically, volunteers were identified from local non-governmental organizations and subsequently trained by Uwezo staff in how to conduct the fieldwork.[4] This has the advantages of keeping costs low (allowing data collection on a much larger scale than otherwise possible), promoting the involvement of local citizens, and ensuring the cultural credibility of the enumerators.

In each primary sampling unit (PSU, typically a village or block of urban households), fieldwork proceeded in four main steps. The first was to survey a single local government primary school that was (randomly) pre-selected by the Uwezo district coordinator. This involved administering a series of questions to the most senior staff member available (ideally the head teacher) and direct observation of pupil and teacher numbers, as well as school conditions according to a simple questionnaire. Second, the chief or administrative head of the PSU was visited. Aside from establishing permission to visit individual households, as well as confirming the validity of the sample frame for the PSU, he or she was asked a series of simple questions about the PSU (e.g., is there access to clean water?). Third, selected households were visited. In each case, the head of the household was asked a short set of simple questions covering the household's general characteristics. Also demographic details of all children in the household were recorded (e.g., age, gender, whether or not attending school etc.). Finally, a series of basic oral literacy and numeracy tests were administered individually to each child in the household. These are discussed in the next sub-section.

Around 350,000 children were surveyed in total for the Uwezo 2 data. However, only the subset of children attending government primary schools are retained in the data analyzed for this study. Children attending a private school (a very small share of observations) are excluded because

[3]Due to frequent changes of administrative boundaries, the administrative divisions from each countries' most recent Population and Housing Census were used. Note that analysis in Uwezo (2012) indicates that aggregate (national) results from the Uwezo 1 and Uwezo 2 surveys are extremely similar.

[4]Fieldwork supervision was provided by experienced volunteers, as well as by a small core team of Uwezo staff. Survey forms were completed on a manual basis in the field, but digital data entry was undertaken centrally in each country by a professional third party firm. In all phases of the fieldwork, training and assistance was provided by ASER. Further documentation is available on the Uwezo website: www.uwezo.net.

we only have information about school inputs from government primary schools. Children not attending any school are excluded as we are interested specifically in the link between constraints to learning that operate within schools (e.g., class sizes). This is not to deny that the former constraints may have an impact on attendance (Hanushek et al. 2006). Nonetheless, these impacts are presumed to be of a second order nature, the first order effect being on learning outcomes for children currently in school.[5] It should also be noted that net enrolment rates are high in East Africa – averaging 85 percent in Kenya and Tanzania, and 91 percent in Uganda for children of compulsory school age (using the present dataset).

Table 1 provides a set of summary statistics from the data used in the remainder of this paper. Panel (a) summarizes the coverage of the retained observations. For each country more than 2,000 individual schools, 30,000 households and 70,000 children are included. Encompassing a total of almost 300,000 children, this represents one of the largest and geographically most comprehensive non-census surveys of its kind, especially for a low-income SSA country. Panel (b) summarizes information from the survey of government primary schools.[6] Pupil (teacher) attendance rates reflect the ratio of the number of pupils (teachers) observed on the day of fieldwork across all classes to the number of formal or officially registered pupils (teachers) according to information provided by the interviewed staff member.

Panel (c) summarizes some of the principal variables that vary at the household level. They include the household's composition and a proxy measure of it's socio-economic status. The latter is calculated from the count of observed physical and human capital assets of the household, as per the Alkire-Foster multidimensional poverty headcount index (Alkire and Foster 2011). Due to limits on the number of candidate variables, only six welfare dimensions are used (given equal weight) – access to electricity, access to piped water, ownership of a phone, ownership of a radio, ownership of a TV, and mother's education. Deprivation with respect to the access/ownership categories is defined as absence of that item in the household; deprivation on the final category is defined as the mother having no formal education. A household thus is defined as 'ultra-poor' if it is simultaneously deprived in all dimensions. If a household is not 'ultra-poor', then it is either defined as 'poor', if is deprived in any four of these dimensions, and 'non-poor' otherwise. Admittedly, these distinctions are somewhat arbitrary; however, this measure is transparent and applies without modification across all countries.

Panel (d) describes a number of key variables that vary at the lowest level of aggregation – i.e.,

[5] Addressing selection-type issues also would significantly complicate the analysis from a technical point of view. For instance, it is a moot point how school inputs might be linked to children that have never attended school.

[6] Section 3.2 looks more closely at the specific class size and composition variables of primary analytical interest.

Table 1: Selected descriptive statistics from the Uwezo dataset

		All	(s.d.)	Kenya	(s.d.)	Tanzania	(s.d.)	Uganda	(s.d.)
(a)	No. districts sampled	318		121		118		79	
	No. schools sampled	9,280		3,449		3,664		2,167	
	No. PSUs sampled	9,642		3,545		3,750		2,347	
	No. households sampled	123,821		44,330		49,947		29,544	
	No. children sampled	252,970		94,171		87,901		70,898	
(b)	Total number of enrolled pupils	611.9	(366.9)	494.8	(296.1)	620.9	(378.8)	709.6	(379.9)
	Attendance rate (%)	81.0	(15.4)	88.8	(9.9)	78.2	(17.0)	77.6	(17.4)
	Teacher attendance rate (%)	80.8	(65.8)	88.5	(27.5)	84.9	(77.4)	67.9	(72.2)
	Number of books for every 10 pupils	27.1	(30.5)	41.3	(25.3)	20.4	(33.5)	22.7	(25.9)
	School has access to clean water (%)	45.0	(24.8)	57.1	(24.5)	27.1	(19.8)	57.5	(24.4)
(c)	Household is poor (%)	47.0	(24.9)	41.4	(24.3)	48.3	(25.0)	50.6	(25.0)
	Household is ultra-poor (%)	7.2	(6.7)	8.6	(7.9)	6.0	(5.6)	7.4	(6.9)
	Number of children in household	3.0	(1.5)	3.0	(1.4)	2.5	(1.2)	3.5	(1.7)
	Number of other household members	4.2	(2.9)	3.7	(2.0)	4.8	(2.9)	4.1	(3.4)
(d)	Age of mother	36.4	(10.2)	35.0	(10.9)	37.9	(10.1)	35.5	(9.4)
	Mother has primary education (%)	65.7	(22.5)	57.5	(24.4)	73.7	(19.4)	62.8	(23.4)
	Mother has secondary education (%)	10.1	(9.1)	16.6	(13.8)	5.0	(4.8)	10.8	(9.6)
	Child is overage-for-grade (%)	44.3	(24.7)	44.6	(24.7)	30.9	(21.4)	61.9	(23.6)
	Child is female (%)	49.5	(25.0)	49.0	(25.0)	50.6	(25.0)	48.6	(25.0)
(e)	Reading test score (%)	57.7	(39.1)	75.0	(31.1)	57.3	(40.3)	42.0	(37.6)
	Comprehension test score (%)	35.6	(22.9)	50.7	(25.0)	36.9	(23.3)	19.6	(15.8)
	Numeracy test score (%)	65.2	(37.1)	79.4	(30.2)	65.0	(37.1)	52.2	(38.2)
	Combined test score (%)	59.8	(35.5)	75.2	(29.7)	59.6	(35.2)	45.5	(34.9)
	Combined test score (std.)	0.0	(100.0)	43.0	(81.8)	-0.8	(99.0)	-39.4	(100.4)

Notes: statistics in panel (a) indicate survey coverage; in the rest of the table statistics are sample means and raw standard deviations (in parentheses); variables in panel (b) are calculated at the PSU level, the remaining variables are calculated over all children of school age attending government primary schools; unless indicated, test scores are rescaled to percentages of the maximum mark; literacy and comprehension tests refer to the predominant national language of instruction at primary school; the combined test score, based on the sum of the literacy, comprehension and numeracy tests, is standardized at the regional-level and multiplied by 100.

Source: author's calculations from the Uwezo data.

children. As some households include children from different parents, these variables include the mother's age and her highest level of completed education. As noted above, the target population is children of school age, up to 16. In Uganda and Kenya, the stipulated age for starting primary school is six. In Tanzania, primary school starts at seven, meaning that six year olds were not assessed on their learning outcomes and are not included in the dataset. This explains the slightly higher average age of mothers in the Tanzanian sample.[7]

Before looking more closely at learning outcomes, as well as the two schooling variables of principal interest, two caveats concerning the data should be highlighted. First, as the Uwezo surveys aimed to address educational issues specific to each of the three countries, the questionnaires and ability tests were developed individually for each country, but based on a common template. As a result, only a subset of questions which were common to each country can be used in cross-country analysis. Moreover, whilst the cognitive tests had the same objectives and format, their precise content differed between countries. This largely reflects slight differences in national curricula evident at the highest levels of tested ability (e.g., as regards expected vocabulary in the story). In the present analysis, these differences have been ameliorated by only including in the test scores the same skill levels assessed in all three countries. Even so, the same mark in Kenya cannot be considered as representing precisely the same absolute level as in Uganda or Tanzania. Indeed, the point of the tests is not to measure underlying cognitive skills in a uniform way (according to some absolute scale), but rather to assess children against the skills they should master by a given grade of primary school in their country of residence.

Second, missing values on specific variables have been imputed for a small number of observations. This was deemed necessary as such missing values appeared to be potentially systematic as opposed to random. Following Wößmann (2005) (also Wößmann 2003), imputation was undertaken using a (random) multiple regression procedure employing 'fundamental' factors such as age, gender and regional location as the explanatory variables. Dummy variables for imputation are included in all regressions reported in Section 5; these capture differences in data

[7] School systems in each country differ in other respects. In Kenya, the duration of primary schooling is the longest at eight years (Standards). According to UNESCO's International Standard Classification of Education (ISCED), the final two years of primary school in Kenya (Standards 7 and 8) correspond to lower secondary schooling on an international scale (ISCED level 2).[8] In contrast, compulsory primary schooling in Uganda and Tanzania is seven years in duration and corresponds to ISCED level 1 only. Aside from age, there are no minimum entrance requirements to enter primary school in either Uganda or Tanzania. Kenya formally encourages at least one unit (year) of pre-primary education to have been completed, although this is not rigorously enforced. Either way, it is widely acknowledged that the vast majority of Ugandan children have no experience of formal education before starting primary school (e.g., see UCRNN 2007).

quality between observations.

3.2 Test scores, class size and peer age composition

The Uwezo tests were designed in each country by local education experts to reflect competencies stipulated in the national curricula at the Standard 2 level. In other words, they test skills that should be achieved by the majority of pupils after two complete years of schooling.[9] Although some primary schools in Kenya and Uganda mainly provide instruction in local languages in the first years of some primary schools, the literacy tests included in the Uwezo surveys refer to the national languages of instruction that pupils are taught throughout primary school and are mandatorily examined in primary school completion examinations – i.e., English and Kiswahili in Kenya and Tanzania; and only English in Uganda.

For the relevant languages, the Uwezo literacy tests evaluated simple reading skills in order of increasing difficulty. Based on pre-prepared test cards, children were asked to: recognize a letter from the alphabet, read a word, read a sentence, and read a paragraph (story). Provided the child is able to read at the story level, she was further asked at least one question to assess whether she comprehended the content of the story. In the numeracy tests, children were asked a set of questions (also from pre-prepared cards) starting with simple arithmetic and either increasing in difficulty, if they were successful, or decreasing in difficulty if not. The numeracy skills assessed in each country thus covered: number recognition, counting, and the performance of basic calculations with numbers of up to two digits (addition, subtraction and multiplication). The numeracy test was administered in the predominant language of instruction (English in Kenya and Uganda; Kiswahili in Tanzania). In each test the child was given a score indicating the maximum skill level achieved.

In comparison to standard education indicators, such as enrolment and grade attainment, the Uwezo test score data provides a direct assessment of acquired cognitive skills. Aggregate results from the Uwezo ability tests are summarized in panel (e) of Table 1, rescaled to percentages. Taking the literacy, comprehension and numeracy tests together, measures of Cronbach's alpha lie between 0.7 and 0.8, calculated individually for each country or for all countries together. As Tavakol and Dennick (2011) explain, this suggests that the different tests measure the same concept or construct (i.e., competence at a Standard 2 level) and, thus, can meaningfully be combined into a single overall mark. Given the diversity of languages spoken across the region,

[9]Exemplars of the tests are available on the Uwezo website.

the literacy and comprehension components of the combined score simply take the language in which the child displays most competence. The combined score is shown in the same table – reported in percentage terms and on a standardized basis. The latter is frequently used in analyses of learning assessments, one of the reasons being that estimated marginal effects represent standard deviation units. The same approach is adopted here and, unless otherwise indicated, the focus is on the standardized combined test score, which takes a mean of zero and standard deviation of 100 (calculated on a pooled basis, across East Africa).

Comparing the test score results across countries, there are marked differences between the three countries, at least when viewed on an unconditional basis. Kenyan pupils outperform Tanzanian pupils by approximately 15 percentage points on each test component and outperform Ugandan pupils by around 20 percentage points. These differences cumulate; the average Ugandan and Kenyan pupil is separated by 80 percent of a standard deviation on the combined test score. Recall that these scores refer only to children enrolled in school; moreover, significant differences are encountered at all ages and therefore are not driven by gaps at specific points in the age distribution.

Turning to the class size and age composition variables, the former is taken from observations of the number of pupils attending a given grade in classes of government primary schools covered by the survey. Note that class size is conceptually distinct from the pupil-teacher ratio. I focus on the former in part because the Uwezo data provides grade-specific information for class sizes; in contrast, the pupil-teacher ratio can only be calculated as a school-wide average. However, particularly in low-income contexts where teacher absenteeism is a chronic problem (Chaudhury et al. 2006), the pupil-teacher ratio may be misleading because it typically includes teachers that are (frequently) absent from class or who undertake administrative duties. Thus, observed class size provides a more reliable measure of the learning conditions which pupils face.

In addition to class size, attention also is directed to so-called overage peer effects. These are defined as the share of a child's classmates who are overage, and where an individual is classified as overage if she is more than one year older than the appropriate age for the grade in which she is enrolled.[10] For example, if the starting age for primary school is the year in which the child turns six, then all children enrolled in Grade 1 aged eight years or older are defined as overage. Likewise, children enrolled in Grade 2 aged nine and over are defined as overage. Children who delay beginning school, those who repeat a grade, as well as others that drop out and restart later,

[10]This definition therefore does not count as overage those children who are were enrolled at the correct age but have an 'early' birthday – i.e., before the survey data was collected.

would all typically be overage. In keeping with the literature on peer effects, the peer overage variable is calculated excluding the individual to whom the observation applies – i.e., only her classmates are counted. Thus, it is denoted as follows:

$$p_{-ik} = \frac{1}{(N_k - 1)} \sum_j o_{jk} \quad \forall j \neq i \tag{1}$$

and where o_{jk} is a dummy variable that takes the value of one, if individual j is overage and N_k denotes the estimated number of pupils in grade k, implictly holding fixed the school and locale. Note that a dummy variable taking the value of one if a child is overage is included in all empirical models (see Section 5), thereby permitting the distinction to be made between direct effects of being overage and indirect effects that work through peer age composition.

Table 2: Summary of class size and overage peer effects

		Country means			Naïve regression	
	Age	KE	TZ	UG	β	Pr.(β=0)
Class size	7	56.4	69.9	98.4	-0.01	0.69
	10	56.8	63.5	87.1	-0.02	0.15
	13	55.5	61.4	76.8	-0.07	0.00
	All	55.7	64.7	83.8	-0.06	0.00
Peers overage (%)	7	32.1	20.4	41.9	-0.02	0.37
	10	43.1	29.9	62.1	-0.09	0.00
	13	51.6	35.9	74.5	-0.10	0.00
	All	44.8	31.6	61.8	-0.06	0.00

Notes: the β coefficients report results from a pooled bivariate regression for pupils of the indicated age, including country fixed effects, where the dependent variable is the combined standardized test score.
Source: author's calculations from the Uwezo data.

Table 2 reports the sample means by country for these two variables at specific ages (and overall). It indicates substantial differences in average class sizes between the three countries. The average Ugandan pupil studies in a class of 84 children, compared to 65 in Tanzania and 56 in Kenya. Class sizes are generally larger for younger children (being 98 for the average Ugandan seven year old), reflecting low rates of survival to completion of primary school and a concentration of late starters and grade repeaters in lower grades. Indeed, the proportion of overage children is striking. With the exception of Tanzanian seven and eight year olds, the average pupil in the region of a given age attends classes where around one in three of her classmates are overage. This phenomenon is most pronounced in Uganda, where on average more than 60 percent of

class peers are overage. This is consistent with concerns that grade repetition is commonplace in Uganda, despite an official policy of automatic promotion (Nishimura et al. 2008, 2009). Tanzania, which in fact permits repetition during grades 1-3 (UNESCO 2012), shows the lowest rates of overage pupils in the region; however, in absolute terms the share remains high.

The final columns of Table 2 indicate the individual associations between the test score outcome and these class size and overage peer effects. Each row corresponds to a single pooled bivariate regression for pupils of the indicated age, including country fixed effects. Thus, the coefficient (column β) indicates the expected marginal change in the test score associated with a one unit increase in the variable of interest. Two results stand out. First, the direction of these naïve marginal effects is consistent with a prior that class size and peer age composition matter – i.e., more classmates or more overage peers are negatively associated with learning outcomes. Second, these coefficients appear to be small in magnitude and are not always statistically significant at conventional levels.

4 Methodology

The associations reported in Table 2 can only be viewed as suggestive – no controls have been included for other covariates and it is hardly conceivable that either of the class size of peer overage variables is uncorrelated with other determinants of learning. This justifies a more detailed multivariate analysis, which attempts to address potential sources of bias that may confound such naïve correlations.

Following a rich existing literature (e.g., Todd and Wolpin 2003), the backbone of the approach adopted to analyse the Uwezo data is an educational production function. In general form, this is implemented as follows:

$$t_i = \lambda_0 + C_i\beta_1 + H_h\beta_2 + p_{-ik}\beta_4 + S_{ks}\beta_5 + G_g\beta_6 + \epsilon_i \tag{2}$$

where indices are nested such that i is the child, h her household, k the grade in which she is enrolled, s the school she attends, and g her geographical location (e.g., administrative district). T is the combined standardized test score which represents a measure of acquired cognitive skills; C contains observed factors that vary at the level of the child (e.g., her age and gender); H contains factors that vary at the level of the household (e.g., household size, socio-economic

status);[11] p is the overage peer effect as per equation (1); S contains schooling inputs that either vary at the level of the grade (e.g., class size) or for the school as a whole; and G captures other contextual conditions, such as the number of primary schools in the locality or whether the village has access to electricty. The final term, ϵ_i, represents a compound residual white noise error term. I assume it can be decomposed into the following normalized (mean-zero) variance components structure:

$$\epsilon_i = \lambda_{1g} + \lambda_{2s} + \lambda_{3k} + \lambda_{4h} + \lambda_{5i} \tag{3}$$

whereby each term on the right-hand side (RHS) corresponds to unobserved effects at alternative levels of aggregation.

Various aspects of this production function merit discussion. First, the model is specified in contemporaneous levels form. That is, past levels of achievement and historical schooling inputs are not included. This is not because they are immaterial. As Todd and Wolpin (2003) clarify, rather stringent assumptions are necessary to obtain consistent estimates of contemporaneous inputs on skills acquisition. However, data constraints mean that alternative specifications of the education production function, such as a value-added model, are not feasible here. Although this constitutes a limitation, a contemporaneous specification is found in a large number of studies (Wößmann 2003; Ammermüller et al. 2005; Aturupane et al. 2011; Glick et al. 2011).

Second, equation (2) is implemented primarily as a reduced form relationship (e.g., Jimenez and Sawada 1999). Endogenous choice or outcome variables, particularly grade attainment, are not retained in the specification as they are viewed as being determined by prior (exogenous) variables already in the model.[12] The rationale for doing so is to capture the total effects associated with marginal changes to the RHS variables, taking into account both the indirect effects on test scores that work through grade attainment, as well as direct effects on test scores conditional on having reached a given grade. *A priori*, it is plausible to expect that a principal channel through which (negative) class and composition effects operate is by impeding grade progression. Controlling for the level of grade attainment would thus only reveal the partial effects on learning outcomes, which may not be sufficiently informative to policy makers.

Nonetheless, structural form models that include grade attainment are also of interest (not least to verify the previous hypothesis). Grade attainment is likely to be associated with historical

[11] To address intra-household distributional effects, household composition is split into two variables – the number of children (aged under 16) and the number of adults. Also, a variable indicating the birthorder of the child is included.

[12] In this sense these endogenous variables can be considered as having been solved out. For elaboration of the distinction between reduced and structural form models see Glewwe et al. (2004).

school inputs and other unobserved factors such as parental support to learning or the child's natural ability. Consequently, inclusion of grade attainment may address sources of bias from some of these omitted factors. Moreover, as Glewwe et al. (2004) note, households might respond to improved school inputs by reducing their own contributions to education. If this is the case, then reduced form models could provide an overly optimistic assessment of the final effects of exogenous changes to school inputs. Crudely speaking, then, reduced form and structural form models respectively indicate something of the upper and lower bounds on the effects of interest. For this reason, structural form results are also reported.

Third, the hierarchical structure of the dataset should be highlighted. As was noted in Section 3.1, individual children are nested in households, which are nested in schools (only one school was surveyed in each PSU), which are nested in administrative districts. The implication is that inclusion of fixed effects at a chosen level of aggregation, corresponding to one of the error terms in equation (3), would absorb all fixed observed and unobserved variables in (2) at the same level or higher. For instance, inclusion of household fixed effects, λ_h, would eliminate all (fixed) variables and parameters uniquely indexed by h, s or g.

Estimation of educational production functions based on cross-sectional observational data raises significant identification challenges (e.g., see Webbink 2005; Hanushek and Rivkin 2006; Glewwe and Kremer 2006), meaning that standard ordinary least squares (OLS) estimates of equation (2) are unlikely to yield unbiased estimates of casual effects. These concerns broadly apply to all variables entering the model, which motivates restricting attention (and further discussion) to identification of the selected class size and peer overage effects. With respect to the former, two principal sources of bias are germane. First, a host of unobserved variables may be correlated with class size. For instance, if additional resources do not flow toward schools with the least resources, class size may be negatively correlated with access to other school resources. Bias from omitted variables also would be consistent with sorting effects, operative at either the household or school levels, such as if low-achieving children are allocated to smaller classes.[13] Second, attenuation bias may be present due to measurement error. The metric of class size employed here is the headcount of pupils in a class (at a given grade) observed on the day of the survey by the enumerators. This could easily diverge from the typical class size experienced by the same pupils due to shocks to pupil or teacher attendance, as well as due to miscounting by enumerators (particularly for larger classes).

[13] A small number of schools in the sample operate separate 'remedial' classes. However, without further information regarding the nature of these classes they are excluded from the analysis.

The overage peer effect potentially suffers from similar sources of bias. For example, poor teacher quality may delay grade attainment, inducing a negative correlation between unobserved metrics of quality and the prevalence of overage children. Delayed grade progression also cumulates over time in the student population, arithmetically producing a positive correlation between a child's observed grade of enrolment and the share of peers that are overage. Thus, in the reduced-form models where grade attainment is omitted, overage peer effects may be confounded with unobserved factors that affect grade attainment. Again, measurement error is a material concern. The peer overage variable is derived from child level observations rather than school level data. On average, a relatively small number of children are observed in each PSU at each grade. If nothing else, this makes the measures noisy.

The above identification challenges are not unique to the present analysis and various solutions have been pursued in comparable studies. One candidate is to employ fixed effects at either the school or household levels. These are feasible here because class size and peer overage variables are not fixed at the school or household levels. For instance, we often observe two or more children in the same household enrolled at different grades. The principal advantage of using fixed effects is that identification is derived from variation within schools or households, thus avoiding confounding from omitted unobserved fixed factors. They also have the advantage of removing sources of bias for all variables retained in the model, not just the variables of principal interest. However, fixed effects are not a panacea. Aside from a possible loss of efficiency, if schools (communities) systematically tend to have either larger or smaller classes (across all grades), then any distinct class size effects that operate at the group level (e.g., across schools or households, not just at the grade level) will be excluded, leading to an underestimate of the overall class size effect (these effects are often labelled compositional effects; e.g., see Algina and Swaminathan 2011; Stock et al. 2011).

A further concern is that the attenuating effects of classical measurement error tend to be exacerbated by inclusion of fixed effects (see Ashenfelter and Krueger 1994; Bound et al. 2001). This implies that addressing one source of bias may come at the expense of inflating others. It should also be noted that when household fixed effects are adopted, in addition to sweeping out many variables of ultimate policy interest (such as its socio-economic status), households observed with only one child become redundant. As typically these are households with older children, this could introduce a source of sample selection bias.

A second approach, which can be complementary to use of fixed effects, is to find instrumental

variables (IVs) for selected variables of interest.[14] To the extent that these IVs are both exogenous and relevant, they should address sources of bias from both unobserved omitted variables and measurement error. This is an attractive proposition; however, in practice it is hard to isolate variables that unambiguously fulfil all conditions for instrument validity. As Stock et al. (2002) comment, in empirical applications there is typically a trade-off between instrument exogeneity and relevance – i.e., instruments with stronger claims to the former frequently tend to display a weaker conditional correlation with the endogenous variable to be instrumented. Moreover, Bound et al. (1995) show that weak instruments significantly magnify bias due to even small violations of the exogeneity requirement. In the absence of experimental conditions (natural or otherwise) that can unambiguously guarantee IVs with the desirable properties, this approach must be used with due caution.

In light of the distinct advantages and disadvantages of the fixed effects and instrumental variables approaches, both procedures are employed. In the former case, fixed effects are specified at different levels of aggregation, enabling coefficient estimates to be compared according to alternative identification assumptions. For the IVs approach, I follow a commonplace approach of constructing instrumental variables from variation occurring at higher (wider) levels of aggregation than that at which the endogenous variables are observed (for examples and discussion see Evans et al. 1992; Card and Krueger 1996; Boozer and Rouse 2001; Ammermüller et al. 2005). The rationale is that whilst variation at a higher level of aggregation reflects general conditions affecting individual pupils, such variation should be independent of idiosyncratic unobserved grade-, household or child-specific variables that represent likely sources of endogeneity bias.

For the class size effect, the preferred instrument is defined as the district level average number of children formally enrolled in classes at each grade. Use of the official number of pupils enrolled in each class (taken from an administrative class list during the school-survey), as opposed to observed attendance on a given day, avoids the concern raised by Pakes (1982) that simple aggregation of observed values of an endogenous variable will fail to address bias arising from measurement error.[15] Focussing on variation at the district level (by grade), also avoids confounding due to fixed unobserved school level factors – i.e., if the instruments had been calculated from class size averages at the school level, then unobserved school level effects

[14] A third approach may be to use hierarchical or mixed modelling approaches, which explicitly take into account the nested structure of the data. A main attraction of using these techniques is to improve estimation efficiency; they do not provide an alternative credible approach to addressing endogeneity such as arising from omitted variables bias. Therefore, they are not considered here.

[15] Enrolment is not chosen as the preferred direct metric of class size due to significant levels of non-attendance among pupils – see Table 1.

could violate the instrument exogeneity requirement (see Boozer and Rouse 2001).

The instrument for the peer overage effect is calculated on a slightly different basis. Averaging across grades at the district level would fail to address unobserved grade-specific effects, such as the tendency for higher grades to contain relatively more overage children. As this is a serious concern, I resort to school-specific averages rather than district level averages by grade. Specifically, the instrument is based on the gap in years between a child's age and the appropriate age for the grade in which she is enrolled. In order to assuage the aforementioned concern that instruments calculated at the school level may be confounded with unobserved school fixed effects, I randomly allocate children from each school to one of five school groups, created at the district level. Subsequently, the mean age gap (in years) for each of these school groups is used as the instrument. Note, as before, that in contrast to using the overage peer effect itself, the age gap variable is used as a basis for calculating the instrument so as to minimize correlation with measurement error in the former variable.

What are the principal threats to the validity of these instruments? First, taken either individually or together, the instruments may only be weakly correlated with the endogenous variables of interest. This cannot be known in advance and must be verified in application. Nevertheless, the limited information maximum likelihood (LIML) instrumental variables (IV) estimator is used, which is considered more robust to weak instruments (Staiger and Stock 1997); also, hypothesis tests that are fully robust to weak instruments are employed (Stock et al. 2002). Second, the instruments may be correlated with unobservable effects, particularly those that operate at either the district (in the case of class size) or school level (in the case of the peer age effect). To address this risk, district level fixed effects are included in all IV regressions.[16] In addition, a full set of covariates specified at the individual, household and school levels is included in the IV specifications. These variables encompass (*inter alia*) the child's age, gender and birthorder; parental ages and level of education; household demographic and socio-economic status (see Section 3.1); overall school size (in log form), the mean teacher attendance rate, books per student and whether or not the head teacher was present when the school was surveyed.

Finally, in order to inspect the validity of the chosen instruments, over-identification tests are reported based on Hansen's J statistic. To do so, at least one additional instrument is required. Following Hoxby (2000), a plausible source of external variation in class sizes is stochastic variation in household size across different localities (controlling for socio-economic conditions etc.). In the present case this is implemented as the number of children born to the mother

[16]School fixed effects cannot be used as they would be collinear with the class composition instruments.

resident in the household averaged over each school group (within each district, as defined above) and age-group bands. As Hoxby and Paserman (1998) note, however, one must be aware that such instrument validity tests are highly sensitive in the context of grouped data, meaning that standard errors must take due account of the (highest) level of aggregation at which grouping takes place. In the present case, therefore, all instrumental variables models adjust for clustering at the district level (by grade); the fixed effects estimates, in contrast, allow for clustering at the community level.

5 Results

The detailed nature of the Uwezo data permits analysis to be undertaken at broad and narrow geographical levels. Region-wide regressions indicate the broad direction and magnitude of estimated relationships. However, more specific (e.g., national) regressions may be of more particular interest because they can help determine whether or not any such average relations are broadly informative across (heterogeneous) sub-populations. Consequently, in presenting the results I focus first on estimates of the international educational production function, in which the three countries are pooled together, meaning that an assumption of slope homogeneity is maintained. Next, I consider country-specific differences. Third, I consider effects for different sub-groups within each country. Each set of results is presented in separate sub-sections.

5.1 International

In order to explore the role of different sets of control variables, Table 3 sequentially adds covariates to a reduced-form specification on the form of equation (2). Throughout, the dependent variable is the combined standardized test score.[17] Column (I) includes only the class size and peer overage effects of interest, as well as child and household characteristics. Column (II) adds a range of other school resources and contextual factors (e.g., access to electricity) observed at the level of the school or community. Columns (III) to (VI) add increasingly more specific (local) fixed effects, starting with three country fixed effects and finishing with over 80,000 household fixed effects.[18] Only selected coefficients and standard errors, clustered at the school

[17] A full list of covariates and detailed empirical results are available from the author on request. Estimates based on the individual reading or numeracy tests are broadly consistent with those based on the combined score.

[18] To avoid an incidental parameters problem the fixed effects specifications are estimated via the 'within' transform. Note that the final specification excludes households with only one observed child.

level, are reported in the table.

A main finding from the fixed effects results is that the coefficient estimates on the class and composition variables are sensitive to modelling choices. The estimates in column (I), which may be thought of as an extreme reduced-form model, broadly conform to the direction and magnitude of the naïve bivariate results reported in Table 2. However, inclusion of geographic effects switches the coefficient on the proportion of overage peer to positive (and significant). Inclusion of school fixed effects approximately doubles the magnitude of both the class size and peer overage coefficients relative to the adjacent district-fixed effects results. One simple interpretation for the differences between these models is that there is significant bias from omitted variables, the net effect of which varies depending on the set of control variables employed. However, measurement errors may be magnified when more specific fixed effects are added.

Three additional points merit comment. First, the goodness of fit of all models is strong, as measured by the R-squared statistics. Even when school controls and geographic fixed effects are excluded (e.g., column I), almost 50 percent of the variation in test scores is accounted for. This compares favourably with results from the literature where it is not uncommon to encounter R-squared statistics of under 30 percent. Second, the direct negative effect of being overage, as opposed to the indirect peer effect due to having overage classmates, is large in both relative and absolute terms. Controlling for other covariates, the predicted test score of an overage-for-grade child is approximately 50 percent of a standard deviation below that of non-overage children of the same age. Of course some gap is expected since non-overage children are, by definition, more advanced in terms of grade attainment. Nonetheless, the magnitude of this gap remains larger than the gains associated with age – the expected test score difference between two children aged one year apart is always less than 30 percent of a standard deviation.[19]

Third, the coefficient estimates for many of the other control variables, including the child being overage-for-grade, vary less dramatically across the columns of Table 3 in contrast to the class size and overage peer effects. In general, coefficient estimates on these other variables decline slightly as additional fixed effects are included. This is consistent with a narrowing of focus onto the 'within' variation, to the exclusion of any compositional (between) effects. It is most noticeable when household fixed effects are introduced, (but also may be due to a change in sample as 35,000 households for which there is only one observed child are now excluded). For

[19] In all specifications age enters as additional fixed effects, thus capturing any non-linearities in expected progression with age.

Table 3: Fixed effects regression estimates (all East Africa)

	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Class size	-0.10***	-0.12***	-0.09***	-0.09***	-0.10***	-0.23***	-0.16***
	-0.01	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Peers overage (%)	-0.03***	0.02*	0.07***	0.08***	0.10***	0.19***	0.17***
	-0.01	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Child is overage-for-grade	-60.28***	-57.77***	-54.70***	-54.30***	-53.65***	-48.56***	-37.56***
	-0.58	(0.58)	(0.58)	(0.58)	(0.57)	(0.55)	(0.85)
Child is female	2.19***	2.23***	2.20***	2.33***	2.38***	2.35***	1.86***
	-0.4	(0.39)	(0.38)	(0.38)	(0.38)	(0.36)	(0.55)
Mother has primary education	3.28***	3.15***	2.64***	3.73***	3.68***	2.84***	4.63
	-0.82	(0.81)	(0.80)	(0.78)	(0.76)	(0.67)	(2.77)
Father has primary education	-1.07	-0.63	1.56	3.19***	3.66***	3.22***	0.08
	-0.89	(0.88)	(0.84)	(0.82)	(0.80)	(0.72)	(3.12)
Mother has secondary education	16.28***	14.15***	13.02***	14.43***	13.02***	10.98***	8.94*
	-1.22	(1.20)	(1.14)	(1.12)	(1.10)	(0.95)	(4.24)
Father has secondary education	12.10***	10.92***	13.23***	15.33***	14.83***	12.83***	3.93
	-1.06	(1.04)	(1.00)	(0.97)	(0.93)	(0.86)	(3.99)
Household is poor	-16.33***	-13.14***	-12.90***	-11.78***	-10.87***	-8.69***	-
	-0.6	(0.57)	(0.56)	(0.54)	(0.53)	(0.49)	
Household is ultra-poor	-23.24***	-18.80***	-18.85***	-17.80***	-16.93***	-14.06***	-
	-1.23	(1.22)	(1.20)	(1.15)	(1.12)	(0.99)	
Included control variables	C, H	C, H, S, G	C, H, S, G	C, H, S, G	C, H, S, G	C, H	C
Grouping variable	None	None	Countries	Regions	Districts	Schools	Households
No. groups	0	0	3	32	318	9,642	88,382
Obs.	252,970	252,970	252,970	252,970	252,970	252,970	217,531
R-squared adj. (total)	0.47	0.48	0.49	0.50	0.51	0.57	0.66
R-squared adj. (within)	0.47	0.48	0.42	0.42	0.42	0.42	0.28
RMSE	72.94	72.2	71.23	70.84	70.18	65.29	58.50

significance: * 0.05 ** 0.01 *** 0.001

Notes: selected coefficients shown; dependent variable is the combined standardized test score; control variables refer to the sets of covariates indicated in equation (2); grouping variables indicates the lowest level at which fixed effects are included, implemented by OLS via the within transformation; sample size in column (VIII) excludes children with no siblings in the dataset; standard errors, shown in parentheses, are robust to clustering at the village level.

Source: author's calculations from the Uwezo data.

Table 4: Fixed effects and IV regression estimates (all East Africa)

Estimator →	FE (I)	FE (II)	IV-LIML (III)	IV-LIML (IV)
Class size	-0.23***	-0.19***	-1.26***	-1.01***
	(0.01)	(0.01)	(0.06)	(0.05)
Peers overage (%)	0.19***	0.09***	-0.57***	-0.40***
	(0.01)	(0.01)	(0.09)	(0.08)
Child is overage-for-grade	-48.56***	-26.95***	-48.24***	-25.40***
	(0.55)	(0.55)	(0.86)	(1.01)
Child is female	2.35***	1.32***	2.25***	1.18**
	(0.36)	(0.35)	(0.43)	(0.39)
Mother has secondary education	10.98***	10.51***	12.24***	11.90***
	(0.95)	(0.90)	(1.23)	(1.12)
Household is poor	-8.69***	-7.91***	-10.10***	-9.22***
	(0.49)	(0.47)	(0.57)	(0.52)
Household is ultra-poor	-14.06***	-12.67***	-16.37***	-14.80***
	(0.99)	(0.95)	(1.18)	(1.09)
Included control variables	C, H	C, H	C, H, S, G	C, H, S, G
Location fixed effects	Schools	Schools	Districts	Districts
Grade fixed effects?	No	Yes	No	Yes
Obs.	252,970	252,970	252,970	252,970
R-squared adj.	0.42	0.47	0.25	0.38
Stock-Wright LM S statistic	-	-	279.66	276.63
Pr. (S)			0.00	0.00
Hansen's J statistic	-	-	1.15	1.27
Pr. (J)			0.28	0.26
Total class size effect at mean	15.6	12.9	85.5	68.5
Total overage peer effect at mean	-19.0	-9.0	57.0	40.0
Overall effect ratio (overage/class)	1.9	1.4	1.2	1.0

significance: * 0.05 ** 0.01 *** 0.001

Notes: selected coefficients shown; dependent variable is the combined standardized test score; control variables refer to the sets of covariates indicated in equation (2); grouping variables indicate the lowest level at which fixed effects are included, implemented by OLS via the within transformation; standard errors, shown in parentheses, are robust to clustering; IV-LIML indicates the instrumental variables LIML estimator; 'class size effect at mean' indicates the total change to the predicted test score for a class of average size from a one child reduction in the class size; 'overage peer effect at mean' similarly is the total change to the predicted test score associated with a one-pupil reduction in the number of overage for grade pupils; 'effect ratio' is the ratio of these effects including, in the numerator, also the direct overage-for-grade effect.

Source: author's calculations from the Uwezo data.

this reason, as well as the fact that household fixed effects sweep out all controls other than those that vary at the child level, the school fixed effects specification is preferred among these fixed effects models.

As discussed previously, fixed effects estimators cannot be relied upon to address bias from measurement error, nor do they address bias from omitted variables that vary at the level of the child, such as factors that affect grade attainment. These concerns are substantiated from the instrumental variables estimates, reported in Table 4. Column (I) of the table replicates the school fixed effects results for the region as a whole; column (II) is the corresponding structural-form model which adds fixed effects for the child's grade of enrolment; column (III) retains district fixed effects, and jointly applies the excluded instruments for the class size and peer effects, as described in Section 4; column (IV) also is the corresponding structural-form model, that includes grade attainment dummy variables.[20] The main result when instrumental variables techniques are deployed is a sharp change in estimated coefficients for the class size variable and the overage peer effects. Column (III) indicates that an exogenous reduction in class size of ten pupils is associated with a 12.6 percent standard deviation increase in expected test scores for an individual student; a ten percentage point fall in the share of overage peers is associated with a 5.7 percent standard deviation test score boost. As expected these coefficient estimates decline in magntiude under the structural-form results, which fits the intuition that class size and peer effects may delay grade attainment. Nonetheless, even when grade of enrolment is included (column IV), significant negative effects continue to be found for both the class size and peer overage variables.

How plausible are these instrumental variables estimates compared to the fixed effects results? The large increase in the magnitude of the class size coefficients is indicative of significant attenuation bias in the latter. As discussed earlier, this is not unreasonable since observed attendance on any given day is likely to be a noisy measure of typical class sizes. Measurement error may also explain some of the difference in the two peer effects estimates. However, to the extent that the IV results are consistent, there may also be some upward bias in the peer overage variable fixed effects estimate, a likely source of which is a positive correlation with factors that drive grade attainment. This view is partially corroborated by the structural-form estimates of the school fixed effects model (column II, Table 4), which yields a smaller coefficient estimate of 0.09 on the peer overage variable. Specification tests reported at the bottom of Table 4, including

[20]The instrumental variables results reported here are consistent with estimates obtained when the class size and peer effects are instrumented individually (separately) rather than jointly. Full details are available on request from the author.

the Stock-Wright LM S statistic which is a weak-instrument robust test of the joint significance of the endogenous variables and the Hansen J test for over-identification, provide comfort that the instruments are valid. Moreover, although there is some loss of efficiency when instrumental variables are employed, the coefficient estimates for other variables remain broadly unaffected. Thus, the LIML-IV approach yields the preferred set of results.

An important point emerges from these results. A direct reading of the parameter estimates from the IV (and fixed effects) results would suggest that the magnitude of the marginal effects associated with class size and overage peers is small, especially relative to the (direct) effects of being overage-for-grade or to other household characteristics. However, this comparison is misleading. Changes in class size and composition typically affect numerous children simultaneously, which is not the case for changes in variables observed at the child (or household) level. For instance, given an average class size of 69 pupils in the sample, if ten pupils were exogenously removed from an average class, then the remaining 59 pupils plausibly gain from being in a smaller class (ignoring the ten pupils removed). Similar arithmetic applies to changes in class composition. Assuming a negative marginal effect from an increase to the share of peers that are overage, if ten overage pupils in a class were exogenously switched for ten non-overage pupils, then all pupils in the new class benefit. The point is that the overall marginal impact of changes to class size or composition must take into account the number of pupils benefiting from such changes, which is not directly apparent from the estimates reported in the main body of Table 4. Also, in addition to indirect effects via changes in peer composition, efforts made to reduce the number of overage children should yield direct gains for these overage pupils, assuming they are in some way able to catch-up. In this sense there may be a double benefit from tackling the prevalence of overage children.

Following the above, the penultimate two rows of Table 4 estimate the combined (total) test score gains for pupils remaining in an average class asscoiated with a one person reduction in the size of the class, and a switch of one pupil from being overage-for-grade to being age-appropriate. From the reduced-form IV results (column III), these effects are substantially larger than the raw marginal effects, being equal to around 85 percent of a standard deviation in the case of the class size change and 57 percent of a standard deviation for the change in share of peers that are overage. The final row of the table reports the ratio of the overage effect to the combined class size effect, where the first component now includes the indirect effect of reducing the share of overage pupils in the class by one pupil (given in the adjacent row) and the direct effect on the selected child (given by the reported overage-for-age marginal effect). These figures must only be considered indicative since endogeneity bias from being overage is not addressed and

no consideration is given to sampling error. Even so, they suggest that the total test score gains from these two hypothetical interventions are comparable in magnitude – the ratio is equal to 1.2 from the reduced-form results, and 1.0 from the structural-form.

The overall effect ratio is informative because it gives a crude indication of the ratio of the costs of the two interventions that should make policy makers indifferent between selecting one or the other. Based on the reduced-form ratio of 1.2, if the cost of reducing class sizes by one pupil is not lower than $1.2^{-1} = 0.77$ of the cost of addressing one overage pupil, then a policy maker should prefer the latter intervention from a cost-benefit perspective. Relative costs will vary between locations; however, in general terms it is reasonable to assume that class size reductions would be the more costly of the two because they invoke additional recurrent costs from employing extra staff. In contrast, addressing the phenomena of grade repetition and late starting, which drive overage-for-grade numbers, is likely to be more feasible via administrative interventions and reorganization of existing resources. As such, one-off investments to deal with existing overage pupils and smaller recurrent costs may be all that is required. If so, then measures to cut the share of overage-for-grade pupils would be substantially more cost-efficient.

5.2 National

Table 5 separately reports results for each country, based on individual fixed effects and the IV regressions. These are broadly consistent with the previous results. In each case the fixed effects estimates for the class size effect appear to be attenuated, while the peer overage effect is upward biased. The overall effect ratio, which indicates something of the comparative effectiveness of addressing class size versus overage-for-grade challenges, also points toward the latter as being a more cost-effective approach. This is particularly pronounced in Uganda where the ratio is greater than 2.0 in both the reduced- and structural-form models.

Notwithstanding these similarities, there are material differences in parameter magnitudes between the three countries. Focussing on the reduced-form IV results (columns Ib, IIb and IIIb), which capture the total effects due to changes in the variables of interest, the class size effect is significantly smaller in Kenya compared to the two other countries. Notably, Kenya has the lowest average class sizes (see Table 2), meaning that a one child reduction in the average class size has an impact on fewer children. This is also reflected in the total class size effect associated with such an intervention, which is equal to 33 percent of a test score standard deviation in Kenya, compared to around 100 percent in Tanzania and Uganda. Overage peer effects are most

substantial in Uganda, which is where the share of overage-for-grade pupils is largest at all ages, in turn yielding a total overage peer effect of 105 percent of a test score standard deviation for the average class.

The country-specific structural-form results are consistent with their reduced-form counterparts and, as expected, the estimated coefficients are smaller. However, changes in parameter estimates when one moves to the structural-form are not the same across countries. These are likely to be driven by differences in the strength of the effect that works through (delayed) grade attainment. In Uganda, for example, both the class size and overage peer effect decline by over one half. Moreover, controlling for grade attainment, there also is no expected direct test score disadvantage in Uganda from being overage-for-grade (column IIIc). One interpretation is that class size and peer effects operate principally through grade attainment. In Kenya and Tanzania there is no statistically significant effect of variations in the share of overage peers once grade attainment is included in the model. Thus, the more overage peers one has, the more likely this will delay a child's own grade attainment.

Finally, although estimates for the remaining covariates are subject to potential bias from omitted variables (and thus cannot be given a causal interpretation), notable differences emerge between countries. On average, girls tend to outperform boys by around 2 percent of a standard deviation (column Ib). This effect is most pronounced in Kenya, while Ugandan girls moderately underperform boys at least when controlling for grade attainment (column IIIc). Parental education and household socio-economic status exert a larger comparative effect in Tanzania – e.g., a Tanzanian child whose mother has secondary education or higher is expected to outperform a comparable child whose mother has no education by around 27 percent of a test score standard deviation; the same gap is less than 10 percent in Kenya and Uganda. This hints at a potential trade-off in dealing with low-achieving children. As noted above, grade retention is associated with large efficiency costs and does not guarantee equal education outcomes (either within grades or overall). However, in the absence of additional support to low-achievers, a policy of automatic promotion may exacerbate differences in competence associated with family background factors. Further consideration of this trade-off goes beyond the scope of this study; nonetheless, it merits deeper attention.

Table 5: National fixed effects and IV regression estimates

Location →	KE			TZ			UG		
Estimator →	FE (Ia)	LIML (Ib)	LIML (Ic)	FE (IIa)	LIML (IIb)	LIML (IIc)	FE (IIIa)	LIML (IIIb)	LIML (IIIc)
Class size	-0.11***	-0.60***	-0.63***	-0.19***	-1.71***	-0.40***	-0.21***	-1.19***	-0.31***
	(0.03)	(0.13)	(0.08)	(0.03)	(0.17)	(0.11)	(0.02)	(0.09)	(0.08)
Peers overage (%)	0.15***	-0.49***	-0.08	0.11***	-0.32*	-0.17	0.30***	-1.05***	-0.41**
	(0.01)	(0.09)	(0.08)	(0.01)	(0.16)	(0.11)	(0.02)	(0.19)	(0.14)
Child is overage-for-grade	-42.49***	-44.64***	-9.64***	-47.50***	-47.95***	-11.09***	-59.72***	-68.50***	-0.86
	(0.73)	(1.34)	(0.96)	(0.98)	(1.47)	(1.36)	(1.18)	(2.80)	(1.24)
Child is female	3.99***	4.01***	2.59***	2.91***	2.03*	0.61	0.48	1.22	-1.82**
	(0.47)	(0.48)	(0.41)	(0.66)	(0.81)	(0.64)	(0.65)	(0.90)	(0.65)
Mother has secondary education	7.50***	5.35***	5.41***	19.20***	28.69***	23.56***	7.66***	7.56**	5.77**
	(1.14)	(1.28)	(1.12)	(2.13)	(2.63)	(1.91)	(1.77)	(2.60)	(1.88)
Household is poor	-6.32***	-7.95***	-6.60***	-10.90***	-13.49***	-11.91***	-7.45***	-7.09***	-4.61***
	(0.64)	(0.64)	(0.55)	(0.85)	(1.02)	(0.81)	(0.92)	(1.24)	(0.84)
Household is ultra-poor	-11.04***	-13.44***	-10.99***	-16.83***	-18.92***	-17.08***	-11.73***	-15.15***	-8.13***
	(1.32)	(1.46)	(1.34)	(1.86)	(2.27)	(1.87)	(1.79)	(2.03)	(1.41)
Included control variables	C, H	C, H, S, G	C, H, S, G	C, H	C, H, S, G	C, H, S, G	C, H	C, H, S, G	C, H, S, G
Location fixed effects	Schools	Districts	Districts	Schools	Districts	Districts	Schools	Districts	Districts
Grade fixed effects?	No	No	Yes	No	No	Yes	No	No	Yes
Obs.	94,171	94,171	94,171	87,901	87,901	87,901	70,898	70,898	70,898
R-squared adj.	0.47	0.42	0.55	0.37	0.11	0.40	0.50	0.27	0.60
Stock-Wright LM S statistic	-	67.17	55.62	-	130.99	18.94	-	156.55	20.78
Pr. (S)		0.00	0.00		0.00	0.00		0.00	0.00
Hansen's J statistic	-	2.29	1.39	-	0.23	0.38	-	0.01	0.01
Pr. (J)		0.13	0.24		0.63	0.54		0.92	0.93
Total class size effect at mean	6.0	32.8	34.5	12.1	108.8	25.5	17.4	98.5	25.7
Total overage peer effect at mean	-15.0	49.0	8.0	-11.0	32.0	17.0	-30.0	105.0	41.0
Overall effect ratio (overage/class)	4.6	2.9	0.5	3.0	0.7	1.1	1.7	1.8	1.6

significance: * 0.05 ** 0.01 *** 0.001

Notes: selected coefficients shown; dependent variable is the combined standardized test score; control variables refer to the sets of covariates indicated in equation (2); grouping variables indicate the lowest level at which fixed effects are included, implemented by OLS via the within transformation; standard errors, shown in parentheses, are robust to clustering; IV-LIML indicates the instrumental variables LIML estimator; 'class size effect at mean' indicates the total change to the predicted test score for a class of average size from a one child reduction in the class size; 'overage peer effect at mean' similarly is the total change to the predicted test score associated with a one-pupil reduction in the number of overage for grade pupils; 'effect ratio' is the ratio of these effects including, in the numerator, also the direct overage-for-grade effect.

Source: author's calculations from the Uwezo data.

Table 6: Summary of reduced-form IV regressions for sub-groups

		Class size		Peers overage		Effect ratios	
		Coeff.	(t-stat.)	Coeff.	(t-stat.)	Indirect	Overall
Kenya	Baseline	-0.60	(-4.65)	-0.49	(-5.50)	1.5	2.9
	Grades 1-3	-1.57	(-4.60)	-0.35	(-1.70)	0.4	1.1
	Grades 4-8	-0.46	(-5.52)	-0.05	(-0.81)	0.2	0.7
	Non-poor	-0.41	(-3.16)	-0.38	(-3.64)	1.7	3.4
	Poor	-0.60	(-4.53)	-0.51	(-4.20)	1.6	3.1
	Ultra-poor	-1.58	(-3.74)	-0.50	(-2.07)	0.6	1.3
Tanzania	Baseline	-1.71	(-9.96)	-0.32	(-2.03)	0.3	0.7
	Grades 1-3	-1.62	(-7.88)	0.06	(0.23)	-0.1	0.3
	Grades 4-7	-0.76	(-4.66)	-0.32	(-2.38)	0.7	1.3
	Non-poor	-1.66	(-7.29)	-0.64	(-2.75)	0.6	1.0
	Poor	-1.71	(-9.23)	-0.03	(-0.16)	0.0	0.5
	Ultra-poor	-1.64	(-3.81)	-0.09	(-0.16)	0.1	0.5
Uganda	Baseline	-1.19	(-13.50)	-1.05	(-5.49)	1.1	1.8
	Grades 1-3	-0.73	(-5.59)	-0.74	(-3.35)	1.1	1.8
	Grades 4-7	-1.37	(-14.12)	-0.65	(-2.79)	0.6	0.9
	Non-poor	-1.11	(-11.16)	-1.00	(-3.51)	1.1	1.9
	Poor	-1.20	(-11.90)	-1.08	(-4.37)	1.1	1.7
	Ultra-poor	-1.47	(-6.96)	-1.25	(-2.47)	1.0	1.6

Notes: rows correspond to individual regressions for the indicated sub-group; dependent variable is the combined standardized test score; only class size and peer overage coefficients shown; models are reduced form IV-LIML, as per column (III), Table 4; 'indirect effect' ratio is the ratio of the 'class size effect at mean' and the 'overage peer effect at mean' as described in Table 4; 'direct effect' adds the direct overage-for-grade effect to the numerator.
Source: author's calculations from the Uwezo data.

5.3 Sub-groups

Results for specific national sub-groups are reported in Table 6. Each row corresponds to a separate reduced-form IV regression for the specified group (e.g., children in grades 1-3 in Kenya). Only estimated coefficients and corresponding t-statistics are reported for the variables of interest, as well as the associated indirect and overall effect ratios. The latter corresponds to the calculations reported in the final row of Table 4, discussed in Section 5.1. The indirect effect ratio reports the ratio of the class size effect to the overage peer effect (excluding the direct overage impact). To calculate both of these ratios, the average class size and share of overage peers for the relevant sub-group is employed, in turn capturing differences in the number of children expected to benefit from class level interventions.

The general pattern of these results indicates no severe departures from the general pattern found at the national level. In other words, the country-wide regression estimates are not wholly misleading for any individual sub-group. But this is not to say the results are homogenous. Two main differences can be highlighted. First, the negative effect due to class sizes and the share of peers that are overage-for-grade varies across grade cohorts. In Kenya, substantial and statistically significant negative effects are found in the earliest grades (1-3). In Tanzania, bigger classes exert a larger negative effect in lower grades; however, overage peer effects are negligible in lower grades but gain (negative) strength and significance beyond grade 4, which is also where the share of overage peers is found to be greater. In Uganda, class size effects are strongest in later grades while the overage peer effect is large and roughly constant across grade cohorts. One explanation for these differences refers to the (mean) test scores differences between countries at each grade. In Kenya, there is strong evidence that children begin primary school comparatively better-prepared to learn – e.g., in grade one the gap between the average Kenyan and Uganda child is around 75 percent of a standard deviation, and is 30 percent of a standard deviation higher than Tanzanian first graders. Thus, the explanation is that class size and composition makes a greater difference in critical periods of learning, which do not occur at the same points in all countries.

The second difference refers to discrepancies between socio-economic groups. With the exception of Tanzania, class size effects are significantly more acute for ultra-poor households. This could reflect a number of factors – for instance, children from ultra-poor households may be more likely go to poorer quality (e.g., small rural) schools; alternatively, poorer households may be less able to compensate for quality differences between schools. At the same time, no clear socio-economic gradient is found for the overage peer effects in either Uganda or Kenya. In

Tanzania, a negative effect appears to be concentrated among non-poor households.

6 Conclusion

The importance of raising the quality of education is widely acknowledged in many developing countries. There is much less consensus, however, as to how quality improvements might be realized. This is partly explained by the lack of information regarding actual learning outcomes and, consequently, a scarcity of solid evidence regarding what determines these outcomes. To help address this gap, the present paper examined the comparative effects on learning outcomes in East Africa of class size and the share of peers (classmates) that are overage-for-grade. The study employed micro-data collected in 2011 from Kenya, mainland Tanzania and Uganda, which incorporates test scores for over 250,000 children aged 6-16. The analysis drew on a standard educational production function framework. However, cognizant of the many potential sources of bias due to omitted variables, including selection effects, alternative methods were used to mitigate these concerns.

The first approach to identification focussed on school fixed effects. While these results indicated a negative effect from larger classes, there were deemed likely to suffer from measurement error as well as bias from non-fixed omitted factors. Consequently, a second and preferred approach employed instrumental variables methods alongside district level fixed effects. Following previous studies, external instruments for the class size and overage peer effects were generated from variation occurring at higher levels of geographic aggregation. Specifically, the instrument for class size used grade-specific of enrolment rates in each district. The instrument for class age composition drew on school-wide averages for randomly chosen groups of schools in each district. In doing so, these instruments also were constructed so as to minimize correlation with measurement error in the underlying variables.

The results from the IV models, while confirming that class size has a negative effect on learning, are substantially larger in magnitude than the fixed effects estimates. Moreover, in contradistinction to the fixed effects estimates, the IV results point to a negative and significant impact of having more overage-for-grade peers. Direct comparison with parameter estimates for other covariates, such as family background factors, suggests that the magnitude of class size and peer composition effects is small. Even so, it was argued that this represents a misleading comparison because marginal changes to classroom conditions typically affect multiple children

simultaneously. These simultaneous impacts were taken into account by calculating a combined or total effect associated with cutting the average class size by one person or with switching one overage-for-grade pupil room for an age-appropriate pupil. These calculations reveal much more considerable effects. For instance, reducing an average class by ten pupils would yield an expected total test score gain of around 73 test score standard deviations for the class as a whole.

The general pattern of results holds for each country individually and for distinct sub-groups within each country. However, notable differences also were revealed. In particular, class size and overage peer effects appear to be much larger in Uganda (based on a reduced-form specification), which is the country with the largest average class sizes and the highest share of overage-for-grade peers. Furthermore, it was suggested that these negative effects work principally via delayed grade attainment in Uganda, pointing to grave difficulties in implementing the official policy of automatic promotion.

Reflecting on the extant literature, this study gives some credence to the view that learning gains may be achieved by relaxing school resource constraints in low-income contexts, especially where these resources are most scarce. However, the present findings also support a view that learning gains do not occur in a vacuum, nor are they unconditional. Peer effects also matter, meaning that changes to the composition of one's classmates can exert a material influence on learning. Combined with clear evidence that overage-for-grade pupils consistently underperform other children, even when controlling for grade attainment, a key conclusion is that the high prevalence of overage-for-grade children demands significant policy attention. This would be in line with a significant (albeit contested) literature which finds grade retention to be an ineffective and inefficient policy to promote human capital development. Indeed, in some situations, addressing this problem may be more cost-effective than reducing class sizes.

At the same time, it must be recalled that a formal policy of automatic promotion has already been adopted in East Africa (at least after the earliest grades). Therefore, it is naïve to suppose that reducing the prevalence of overage-for-grade pupils can be achieved quickly or easily. Rather than reject the 'cure' (automatic promotion) in favour of the original 'disease' (grade retention), however, a key policy challenge is to make the cure more effective. To forestall a widening of educational inequalities that may be inimical to the public acceptance of an automatic promotion policy, part of this is likely to require targeting additional support to low-achieving pupils and schools. Addressing large class sizes and the prevalence of overage-for-grade children therefore cannot be seen in binary or exclusive terms. The priority is to enhance learning outcomes, and rather than focus on individual 'solutions', incentives must be fundamentally strengthened to

attain this goal.

References

Algina, J. and Swaminathan, H. (2011). Centering in two-level nested designs. In J.J. Hox and j. Kyle Roberts (Eds.), *Handbook of advanced multilevel analysis*, chapter 15, pp. 285–312. Ney York: Routledge.

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7):476–487.

Ammermüller, A., Heijke, H. and Wößmann, L. (2005). Schooling quality in Eastern Europe: Educational production during transition. *Economics of Education Review*, 24(5):579–599.

Ashenfelter, O. and Krueger, A. (1994). Estimates of the Economic Return to Schooling from a New Sample of Twins. *The American Economic Review*, 84(5):1157–1173.

Aturupane, H., Glewwe, P. and Wisniewski, S. (2011). The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools. *Education Economics*, iFirst Article:1–37.

Betts, J.R. and Shkolnik, J.L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, 19(1):21–26.

Black, S.E., Devereux, P.J. and Salvanes, K.G. (2011). Too Young to Leave the Nest? The Effects of School Starting Age. *The Review of Economics and Statistics*, 93(2):455–467.

Bold, T., Kimenyi, M., Mwabu, G. and Welcome, P. (2012). Interventions & Institutions Experimental Evidence on Scaling up Education Reforms in Kenya. URL www.iies.su.se/polopoly_fs/1.101632.1348137980!/menu/standard/file/2012-08-06%20Kenya%20RCT.pdf.

Boozer, M. and Rouse, C. (2001). Intraschool Variation in Class Size: Patterns and Implications. *Journal of Urban Economics*, 50(1):163–189.

Bound, J., Brown, C. and Mathiowetz, N. (2001). Measurement error in survey data. *Handbook of Econometrics*, 5:3705–3843.

Bound, J., Jaeger, D. and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Brophy, J. (2006). Grade repetition. Education policy series 6, UNESCO: International Institute for Educational Planning. URL http://unesdoc.unesco.org/images/0015/001520/152038e.pdf.

Card, D. and Krueger, A.B. (1996). School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina. *The Journal of Economic Perspectives*, 10(4):31–50.

Case, A. and Deaton, A. (1999). School inputs and educational outcomes in South Africa. *The Quarterly Journal of Economics*, 114(3):1047–1084.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K. and Rogers, F.H. (2006). Missing in action: teacher and health worker absence in developing countries. *The Journal of Economic Perspectives*, 20(1):91–116.

Chingos, M.M. (2012). Class Size and Student Outcomes: Research and Policy Implications. *Journal of Policy Analysis and Management*, forthcoming. doi:10.1002/pam.21677.

Cho, H., Glewwe, P. and Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review*, 31(3):77–95.

Deming, D. and Dynarski, S. (2008). The Lengthening of Childhood. *The Journal of Economic Perspectives*, 22(3):71–92.

Duflo, E., Dupas, P. and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *The American Economic Review*, 101:1739–1774.

—— (2012). School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools. Working Paper 17939, National Bureau of Economic Research. URL www.nber.org/papers/w17939.

Epple, D. and Romano, R. (2011). Peer effects in education: A survey of the theory and evidence. *Handbook of Social Economics*, 1(11):1053–1163.

Evans, W.N., Oates, W.E. and Schwab, R.M. (1992). Measuring Peer Group Effects: A Study of Teenage Behavior. *The Journal of Political Economy*, 100(5):966–991.

Fehrler, S., Michaelowa, K. and Wechtler, A. (2009). The Effectiveness of Inputs in Primary Education: Insights from Recent Student Surveys for sub-Saharan Africa. *The Journal of Development Studies*, 45(9):1545–1578.

Figlio, D. (1999). Functional form and the estimated effects of school resources. *Economics of Education Review*, 18(2):241–252.

Fuller, B. (1986). Is primary school quality eroding in the Third World? *Comparative Education Review*, 30(4):491–507.

Glewwe, P. and Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, 2:945–1017.

Glewwe, P., Kremer, M., Moulin, S. and Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics*, 74(1):251–268.

Glick, P., Randrianarisoa, J. and Sahn, D. (2011). Family background, school characteristics, and children's cognitive achievement in Madagascar. *Education Economics*, 19(4):363–396.

Glick, P. and Sahn, D. (2009). Cognitive skills among children in Senegal: Disentangling the roles of schooling and family background. *Economics of Education Review*, 28(2):178–188.

Hanushek, E., Kain, J., Markman, J. and Rivkin, S. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5):527–544.

Hanushek, E.A., Lavy, V. and Hitomi, K. (2006). Do Students Care about School Quality? Determinants of Dropout Behavior in Developing Countries. NBER Working Papers 12737, National Bureau of Economic Research, Inc.

Hanushek, E. and Rivkin, S. (2006). Teacher quality. *Handbook of the Economics of Education*, 2:1051–1078.

Hanushek, E.A. and Wößmann, L. (2011). The Economics of International Differences in Educational Achievement. In E.A. Hanushek, S. Machin and L. Wößmann (Eds.), *Handbooks in Economics*, volume 3, pp. 89–200. The Netherlands: North-Holland.

Hattie, J.A. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, 37(5):449–481.

Heyneman, S.P. and Loxley, W.A. (1983). The Effect of Primary-School Quality on Academic Achievement Across Twenty-nine High- and Low-Income Countries. *American Journal of Sociology*, 88(6):1162–1194.

Hoxby, C.M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, 115(4):1239–1285.

Hoxby, C. and Paserman, M.D. (1998). Overidentification Tests with Grouped Data. NBER Technical Working Papers 0223, National Bureau of Economic Research, Inc.

Jimenez, E. and Sawada, Y. (1999). Do community-managed schools work? An evaluation of El Salvador's EDUCO program. *The World Bank Economic Review*, 13(3):415–441.

Kremer, M. and Holla, A. (2009). Improving Education in the Developing World: What Have We Learned from Randomized Evaluations? *Annu. Rev. Econ.*, 1(1):513–542.

Lavy, V., Paserman, M.D. and Schlosser, A. (2012). Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *The Economic Journal*, 122(559):208–237. doi:10.1111/j.1468-0297.2011.02463.x.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics*, 94(2):596–606.

Michaelowa, K. (2001). Primary education quality in francophone Sub-Saharan Africa: Determinants of learning achievement and efficiency considerations. *World Development*, 29(10):1699–1716.

—— (2003). Determinants of primary education quality: What can we learn from PASEC for francophone Sub-Saharan Africa. Background paper for the adea study: 'the challenge of learning: improving the quality of basic education in sub-saharan africa', Association for the Development of Education in Africa (ADEA), Paris.

Motala, S. (1995). Surviving the system: a critical appraisal of some conventional wisdoms in primary education in South Africa. *Comparative Education*, 31(2):161–180. doi:10.1080/03050069529092.

Ndaruhutse, S. (2008). Grade repetition in primary schools in Sub-Saharan Africa: an evidence base or change. CfBT Research Report, CfBT Education Trust. URL www.cfbt.com/evidenceforeducation/pdf/Grade%20Repetition_FINAL_8FEB08.pdf.

Nishimura, M., Ogawa, K., Sifuna, D.N., Chimombo, J., Kunje, D., Ampiah, J.G., Byamugisha, A., Sawamura, N. and Yamada, S. (2009). A comparative analysis of universal primary education policy in Ghana, Kenya, Malawi, and Uganda. *Journal of International Cooperation in Education*, 12(1):143–158.

Nishimura, M., Yamano, T. and Sasaoka, Y. (2008). Impacts of the universal primary education policy on educational attainment and private costs in rural Uganda. *International Journal of Educational Development*, 28(2):161–175.

Okuni, A. (2003). Quantity-quality trade-offs after UPE: prospects and challenges of universal access in Uganda. *Norrag News*, 32:32–37. URL www.norrag.org/issues/download/NN/32/en.

Pakes, A. (1982). On the asymptotic bias of Wald-type estimators of a straight line when both variables are subject to error. *International Economic Review*, 23(2):491–497.

Republic of Uganda (2010). National Development Plan (2010/11–2014/15). Technical report, National Planning Authority. URL http://npa.ug/docs/NDP2.pdf.

Roderick, M., Nagaoka, J. and Allensworth, E. (2005). Is the Glass Half Full or Mostly Empty? Ending Social Promotion in Chicago. *Yearbook of the National Society for the Study of Education*, 104(2):223–259. doi:10.1111/j.1744-7984.2005.00032.x.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? *Handbook of the Economics of Education*, 3:249–277.

Staiger, D. and Stock, J.H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, C., Ejstrud, B., Vinther-Larsen, M., Schlattmann, P., Curtis, T., Grønbæk, M. and Bloomfield, K. (2011). Effects of school district factors on alcohol consumption: results of a multi-level analysis among Danish adolescents. *The European Journal of Public Health*, 21(4):449–455.

Stock, J.H., Wright, J.H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2:53–55.

Todd, P. and Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33.

UCRNN (2007). Hope amidst Obstacles: The State of Nursery Education in Uganda. Policy report, Uganda Child Rights NGO Network (UCRNN). URL www.ucrnn.org/resources/researchpapers/Nursery%20report%20Print%20version.pdf.

UNESCO (2011). *Global Education Digest 2011: Comparing Education Statistics Across The World*. Quebec: UNESCO Institute for Statistics.

—— (2012). *World Data on Education 2010/11*. UNESCO: International Bureau of Education, vii edition. URL http://www.ibe.unesco.org/en/services/online-materials/world-data-on-education/seventh-edition-2010-11.html.

Uwezo (2012). Are our Children Learning? Literacy and Numeracy across East Africa. Policy report, Uwezo. URL www.uwezo.net/wp-content/uploads/2012/08/RO_2012_UwezoEastAfricaREport.pdf.

Webbink, D. (2005). Causal effects in education. *Journal of Economic Surveys*, 19(4):535–560.

Wells, R. (2009). Gender and age-appropriate enrolment in Uganda. *International Journal of Educational Research*, 48(1):40–50.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics*, 65(2):117–170.

—— (2005). Educational production in Europe. *Economic Policy*, 20(43):445–504.