Estimating effects of school quality using multiple proxies

Pedro Bernal a, Nikolas Mittag b,⁎, Javaeria A. Qureshi c

a Inter-American Development Bank, 1300 New York Ave. NW, Washington, D.C. 20577, USA
b CERGE-EI, Politických vězí 7, 111 21 Praha 1, Czech Republic
c Department of Economics, University of Illinois at Chicago, 601 South Morgan Street, 716 University Hall, Chicago, IL 60607, USA

HIGHLIGHTS

• We estimate a proxy variable model to identify the effect of latent school quality.
• We find significant effects of teaching and resource quality on student achievement.
• The effects are relatively small, but imply sizable life-time increases in earnings.
• Conventional estimates understate the effect of school quality by about 50%.
• Measurement error may reconcile the ambiguous evidence on effects of school quality.

ABSTRACT

The recent literature on school quality has shown that the school a child attends has significant effects on achievement. However, the literature relating different school characteristics to student achievement has produced mixed results, particularly when using student-level data. Using data from the ECLS-K and a proxy variable model that addresses the problem of measuring school quality, we show that significant effects of teaching and resource quality can be detected from student-level data. We find a significant, positive relationship between school quality and student achievement if school characteristics such as class size and teachers' schooling are treated as noisy measures of school quality. However, this effect is not detected when using models which do not account for measurement error in school quality. Our results suggest that conventional approaches underestimate the effect of school quality by about 50%.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

JEL classification:
C38
I21
I20
J24

Keywords:
School quality
Education
Skill formation
Proxy variables
Measurement error

1. Introduction

The impact of school quality on student achievement has been heavily debated since the publication of the Coleman Report, which found relatively small effects of differences in the measured attributes of schools on student outcomes (Coleman et al. (1966)). On the one hand, the importance attached to school choice and resources invested by parents and policy makers in schools suggests school quality plays an important role in child development. This is supported by evidence of the importance of good teachers (e.g. Rockoff (2004); Rivkin et al. (2005); Jackson (2013); Chetty et al. (2014)) and school level comparisons using quasi-random variation in school assignment which show significant effects on student outcomes (e.g. Hastings and Weinstein (2007); Pop-Eleches and Urquiola (2013)). On the other hand, several similar school-level studies fail to find an impact on student achievement (e.g. Clark (2010); Cullen et al. (2005)). The evidence from the vast literature analyzing the effect of school quality using measures such as class size, teacher characteristics, or expenditure per capita on student outcomes is also mixed. For example, Hanushek’s (2003) review finds that among 276 estimates of the effect of student-teacher ratio on student performance, 14% of studies found positive and statistically significant effects while another 14% found significant, negative effects.

This paper seeks to re-investigate the link between school attributes, school quality and test score achievement. We argue that school attributes – such as teacher’s schooling and class size – that are often used to explain student achievement measure school quality, which is
unobserved, with considerable error. To the extent that these measures proxy latent school quality with error, existing estimates of school quality may exhibit substantial biases. Consequently, the fact that the past literature does not consistently detect a significant impact of school quality may not be due to the absence of a relation between school quality and student outcomes. Understanding the role of school quality in determining student achievement is important given the significant returns to better test scores (e.g. school attainment: Currie and Thomas (1999); Murnane et al. (2000) and wages: Murnane et al. (1995)). It is also important to account for the role of school quality to avoid biases in studies of skill formation (e.g. Cunha and Heckman (2008); Cunha et al. (2010); Todd and Wolpin (2003)). For example, ignoring the role of school quality is likely to lead to overestimates of the own-productivity of skills.

This study estimates the effect of school quality on student achievement using multiple measures of school characteristics in an extension of the proxy variable model developed by Black and Smith (2006). We find that two latent dimensions of school quality – teaching and resource quality – affect achievement. We also develop a test of the validity of the model by deriving the results one would expect to obtain from a model that does not account for the measurement error, and comparing the implied results with the actual results obtained from estimating such a model. The results of this exercise and several other analyses we perform provide strong evidence in favor of the latent quality approach, which accounts for the presence of measurement error.

We use the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) data to estimate the effect of school quality, since it contains information on student, parent, teacher and school characteristics. We provide evidence that suggests the rich set of student-level controls we use is sufficient to account for the endogeneity of school quality. The ECLS-K provides us with several measures of school characteristics to use as proxies for school quality. We use commonly used measures of teachers’ schooling, certifications and related college courses as proxies for teaching quality, and measures such as class size, access to instructional computers and specialized staff as proxies for resource quality.

Using the proxy variable model, we find significant, positive impacts of school teaching and resource quality on student achievement.1 While we do not detect an effect of school quality on math achievement, we find a significant effect on reading achievement that is small but important. An increase of one standard deviation in both teaching quality and resource quality is associated with an improvement of 0.77 in standardized reading test scores between the spring of kindergarten and the spring of first grade. This effect on reading achievement corresponds to roughly 25% of the additional widening of the black-white reading test score gap that takes place between the fall of kindergarten and spring of first grade.

Since the effect of school quality is small and the proxies are noisy, we find that ignoring measurement error in school quality leads to substantial bias. We show that models that do not account for this measurement error tend to conceal the positive impact of school quality on student achievement: they yield estimates that are 50% attenuated, on average, relative to the estimates we find using our proxy variable model. The consequences of measurement error for estimation in combination with the small effect of school quality on student achievement can explain the conflicting evidence from past studies. Our proxy variable model detects a significant effect of school quality in individual-level data, suggesting that taking measurement error into account may also reconcile the discrepancy in evidence from aggregate or school level studies, which tend to find significant effects, with that from individual level studies where such effects have been harder to detect.2

The rest of the paper proceeds as follows: Section 2 discusses how we define school quality and how it relates to previous definitions. Section 3 formulates the education production function, and Section 4 describes our estimation strategy. Sections 5 and 6 describe the ECLS-K data and our choice of proxies for school quality. Section 7 presents our results, tests of the validity of our model and discusses policy implications, and Section 8 concludes.

2 Defining school quality

Many previous studies have analyzed the effect of school quality using measures such as class size, teacher characteristics, or expenditures per capita to infer their impact on student outcomes (e.g. Angrist and Lavy (1999); Chetty et al. (2011); Dynarski et al. (2013); Goldhaber and Brewer, (2000); Hanushek (1997); Rivkin et al. (2005)). These studies consider these variables to be direct inputs in the achievement production function. Since it assumes a direct causal relationship between the input variables and outcomes, this approach does not need the concept of school quality. While some of these studies find a significant effect, others do not (see e.g. Hanushek (2003) for an overview). We argue that the failure to detect an impact is not due to the absence of a relation between school quality and student outcomes, but that a positive and significant relationship is detected when these variables are treated as noisy measures of school quality.

Recent studies comparing the outcomes of students across different schools (e.g. Pop-Eleches and Urquiola (2013); Hastings and Weinstein (2007)) and teachers (e.g. Rivkin et al. (2005); Rockoff (2004); Jackson (2013); Chetty et al. (2014)) using fixed effects approaches and quasi-experimental methods have provided credible evidence that schools and teachers matter. There is still some disagreement on the existence and size of these impacts, see e.g. Clark (2010) or Cullen et al. (2005) for studies that do not find an impact. However, these studies leave some important questions unanswered. Many of the studies that establish a link between schools and student outcomes make a binary comparison between school types, showing that attending private schools, charter schools or more selective schools leads to better student outcomes. These comparisons leave the measurement of school quality implicit, which has downsides. First, it makes it hard to identify the mechanisms that lead to improved student outcomes because the schools differ in many ways and it remains unclear which differences matter for student learning (see e.g. Angrist et al. (2013)). Second, the magnitude of the effects is difficult to interpret. Since school quality is not explicitly measured, it remains unclear whether moving from, for example, a less selective to a more selective school constitutes a small or large change in school quality. Tying the improvements in outcomes to an interpretable metric of school quality is important to identify good schools and to answer policy questions, such as what the likely effects of higher investments in schools or transferring between schools that were not explicitly studied would be.

Similar arguments apply to the literature on teacher fixed effects, which demonstrates that teacher quality is an important determinant of student achievement but is uninformative about how to identify a good teacher. Rockoff et al. (2011) attempt to address this issue by aggregating noisy measures to scores that predict teacher quality before hiring them. We take a similar approach to the problem of school quality. We argue that the variables that are commonly used as measures of school quality – such as class size and teacher education – can be considered noisy proxies for school quality. School quality produces achievement, but is latent and unobserved. The essence of our method is to use several of these noisy proxies to extract the signal they contain about school quality, which allows us to detect an impact of teaching2

1 Teaching and resources are not necessarily the only dimensions that matter, but they are the ones we are able to detect in our data. Examples of other possible dimensions include parental involvement and peer effects (see Smith and Stange (2015), for the case of college quality). We did not find an impact of these dimensions (possibly due to a lack of power or good proxies), so we leave this issue for future research.

2 Betts (1995) and Hanushek et al. (1996) point out this pattern of results by aggregation level in studies.

3 It is important to distinguish teaching quality from the effect of a particular teacher in the literature on teacher fixed effects. Teaching quality as we define it is a school level characteristic and not tied to particular teachers and classrooms. See Section 6 for further discussion.
and resource quality on student test scores. We assume that the proxy variables like class size are related to school quality, but do not impose the requirement that they produce achievement or school quality directly. Rather, we assume that school quality is something unobserved about a school that causes more or less achievement to be produced and that this unobserved characteristic of a school is systematically but not necessarily causally related to these variables. Thus, our approach does not require them to cause school quality.

While estimating the impacts of individual school factors such as class size or teacher qualifications on achievement can yield policy relevant evidence on specific changes that can be implemented in schools, our method can better address the question of whether school quality matters overall. While it has the relative disadvantage that it does not necessarily inform us about which inputs into school quality matter most for student achievement, we show that ignoring the proxy variable formulation leads to considerable bias, so that taking measurement error into account is necessary to obtain credible estimates of the effect of school quality. Our method also has the advantage of making the effects of school quality more interpretable by establishing a metric that can be applied to schools that were not subjects of the initial study.

3. The impact of school quality on student achievement

We estimate an education production function that relates a student’s test score to the quality of the school the student attends and a set of controls aimed at removing confounding factors such as family background and ability. In particular, we assume that the stock of achievement at time \( t \), as measured by a test score \( \theta_t \), is produced according to a production function such as

\[
\theta_t = f(Q_t, \theta_{t-1}, O)
\]

where \( Q_t \) is school quality, \( \theta_{t-1} \) is achievement in the previous period and \( O \) includes other factors that may influence achievement such as parental investments. Eq. (1) is a standard model of human capital formation (see e.g. Boardman and Murnane (1979) and Todd and Wolpin (2003)). Achievement can differ by subject (such as math, reading, etc.), and may be produced according to different functions, so \( \theta_t \) is a vector. All inputs could be vectors, e.g. to include past or cumulative values. There could be multiple dimensions of school quality, such as the quality of resources, quality of teaching, quality of peers etc., so \( Q_t \) may not be a single index but a vector. The impact of school quality can also be different for different kinds of achievement in \( \theta_t \).

Contrary to some of the studies discussed in Section 2, we consider \( Q_t \) to be unobserved and the variables they consider direct inputs to be noisy proxies for latent school quality. As Black and Smith (2006) discuss in the context of college quality, both approaches are conceptually valid and yield different parameters of interest. Conceptualizing the variables we use as proxies for \( Q_t \) as direct inputs is a restricted case of the model we propose, where the measures contain no error and there are as many dimensions of school quality as there are inputs. In Section 7, we perform tests that provide strong evidence for the generalized model which accounts for measurement error. We also empirically evaluate the restricted version of the model in our sample, and show that the results and conclusions are substantially different from those obtained from the generalized model.

We do not assume a causal relationship between the proxies and school quality and only require the weaker assumption that these measures send a noisy signal about the quality of the school. Some of them may send this signal because they are producing school quality, and this is consistent with, but not necessary for our generalized approach. While our approach does not tell us how one can change school quality by manipulating the inputs, the method we propose could be extended to estimate production functions in order to identify a causal relationship between the proxies and school quality and examine how to change school quality. However, this requires stronger assumptions, so this study focuses on the first step question of clarifying the relationship between school quality and achievement. Since our approach identifies variables that are related to school quality it still provides suggestive evidence on likely inputs into the production of school quality.

The mixed evidence on the role of school quality is generally consistent with the approach we suggest. While studies that compare students across different types of schools (e.g. Pop-Eleches and Urquiola (2013); Hastings and Weinstein (2007)) and studies using data at a higher level of aggregation such as the state tend to find significant, positive effects of school quality, studies relating individual- or school-level outcomes to school inputs find effects that are less significant and much smaller (Hanushek et al. (1996); Betts (1995)). If variables commonly used as direct inputs are imperfect measures of school quality, it should be hard to discern the impact of school quality in micro data because the measurement error will lead to biased and most likely attenuated coefficients. On the other hand, one should be more likely to find an impact of school quality in aggregate data, because aggregating the data often reduces measurement error. Thus, the presence of measurement error in the measures of school quality can theoretically help to explain why the relationship between school quality and achievement tends to be discernible in aggregate data but less so in individual data.

We also expect the bias from measurement error to tend to mute the effect of school quality if one directly links school attributes to student outcomes without correcting for the measurement error (see Black and Smith (2006) for a discussion). In Section 7.1, we discuss the implications of measurement error for models that do not take it into account, and use this to provide evidence in support of our latent quality approach.

4. The econometric model

Several recent papers highlight the importance of extracting a signal from noisy proxies in empirical work related to ours (e.g. Cunha and Heckman (2007); Rockoff et al. (2011)). Most closely related, Black and Smith (2006) model measurement error in college quality to estimate the returns to college quality. We use the same approach of using multiple proxies to estimate the impact of latent school quality, \( Q_t \), on student achievement. Unlike Black and Smith (2006) who estimate the effect of unidimensional college quality, we allow for two types of school quality, teaching and resource quality. To do so, we extend their GMM estimator to multiple latent variables. We combine this proxy variable model with a rich set of controls to address the non-random selection of students into schools. We first describe how to estimate the education production function if school quality were observable in Section 4.1 and then extend this strategy to unobservable school quality in Section 4.2. We discuss the problem of selection in Section 4.3.

---

4 The effect of teacher education or class size is of interest to policy makers and school administrators who have to choose which factors to focus on when cutting or augmenting a college budget, while the effect of school quality is more interesting when budgetary allocations within a school are not the primary policy issue of interest.

5 If school quality is produced by a Cobb-Douglas production function, the inputs satisfy the assumptions of our model.

6 Aggregation can reduce attenuation bias due to measurement error if it raises the signal-to-noise ratio (Hanushek et al. (1996)). They show that replacing the microdata with the group average is a variant of instrumental variables in which the group average is used as the instrument, and can, therefore, eliminate attenuation bias due to classical measurement error. Hanushek et al. (1996) present a theoretical model to demonstrate this expected effect of aggregation on measurement error bias, but they do not find empirical support for it in their application.
4.1. The education production function

We estimate the production function separately for achievement in math and reading. The impact of school quality on achievement is difficult to identify due to the endogenous sorting of children into schools. We do not attempt to specify the other inputs or estimate their impact but only try to control for them in a way that allows us to identify the effect of Qit. Thus, we try to find a vector of controls, X, for each type of achievement, such that conditional on X and the lagged test score θt-1, the remaining unobserved factors in 0 are uncorrelated with Qit. The education production function we estimate takes the following form:

\[ θ_t = Q_{it} \alpha + Xβ + θ_{t-1} γ + ε_t \]  

(2)

Except for the fact that school quality is unobserved, Eq. (2) is a linear regression model. The maintained assumption is that the controls in X and the lagged test score capture the impact of all inputs that are correlated with current school quality:

\[ E(ε_t | Q_{it}, X, θ_{t-1}) = 0 \]  

(A1)

Thus, we discuss selection on observables only and linearity. We discuss the conditions under which selection may cause bias in Section 4.3 and provide empirical evidence in Section 7.1 that the rich controls in the ECLS-K data make selection on observables plausible. To justify linearity in the controls, X includes higher order terms of key variables and Section 7.1 shows that our results are robust to changes in the controls.

However, we do not include any non-linear terms of Qit for two reasons. First, neither Qit nor θ have a fixed scale. Thus, a transformation that makes θ linear in Qit exists under fairly innocuous assumptions. Second, we find no empirical evidence of a non-linear relationship. Nonetheless, a non-linear relation is not implausible and our sample is not large enough to rule it out or to estimate it precisely. The model is still identified with non-linear terms, so relaxing this assumption would be an interesting extension with other data. However, it is unlikely that test scores are non-monotonic in school quality, so even if there are non-linearities, our results still present a linear approximation and establish a positive impact of school quality on achievement.

4.2. Inferring the impact of school quality from noisy proxies

Eq. (2) cannot be estimated directly, since Qit is not observed. However, the ECLS-K data include several proxies, qit, such as class size and measures of the quality of the teachers such as their years of education and certifications. Estimating Eq. (2) using proxies instead of Qit yields biased coefficients due to measurement error. To obtain consistent estimates, we use an extension of the proxy variable model proposed by Black and Smith (2006). We assume that the proxies can be represented in the following linear projection form:

\[ q_{it} = δ_{0i} + Q_{it} δ_{1i} + Xδ_{2i} + θ_{t-1} δ_{3i} + η_{it} \]  

(A2)

where qit represents the ith proxy for school quality. The K × 1 vector Qit contains the K different types of school quality, so δ1i is a vector. We assume that conditional on X and θt-1, the error terms in A2 are uncorrelated with both the other error terms and Qit, i.e.:

\[ Cov(η_{it}, q_{jt}) = 0 \]  

(A3)
proxies with the latent variable, while the attenuation of coefficients in a factor model is only reduced to zero as these correlations approach one. This can be an important advantage when the correlations (i.e. the parameters in A) are small. Secondly, contrary to our approach, factor models require prediction of factor scores and inclusion of them in a regression, both of which can introduce bias (see e.g. Grice (2001); Croon (2002) and Skrondal and Laake (2001)). Finally, GMM easily extends to clustered errors and avoids common distributional assumptions of other models.

4.3. School quality and selection into schools

A problem affecting all studies that estimate the impact of school quality on student outcomes is the non-random selection of students into schools. As a first step to address this issue, we estimate the impact of school quality in first grade only as opposed to estimating its effect in later grades, thereby limiting the influence that past inputs and past school quality could have on test scores. Estimating a fully dynamic model is feasible, but it would be more susceptible to misspecification, and our restricted specification strengthens the plausibility of A1 by reducing concerns about attrition and selection. Our model relates a child’s first grade spring test score to a rich set of controls, and his/her kindergarten spring test score. In estimates of education production functions, lagged achievement is often assumed to be a sufficient statistic for latent ability and the unobserved history of inputs until that period (Todd and Wolpin (2003)). Our specification should, therefore, control for any sorting into schools based on the controls in X, time-invariant inputs and child ability as measured by the lagged score. Consequently, only selection on changes in inputs that happen after the kindergarten score, i.e. between spring of kindergarten and first grade, could violate A1 if these changes are orthogonal to our controls, but not the first grade test score and qx. This may be the case if parents alter their investments between kindergarten and first grade differentially across schools of different quality. This requires strong conditions on parental behaviour, however, and studies such as Chetty et al. (2014) find little evidence of selection after conditioning on a lagged test score with a much smaller set of controls. Similarly, we find that few controls actually predict first grade scores after conditioning on the lagged test score.

If many children switch schools between kindergarten and first grade, controlling for the kindergarten score may not be sufficient to control for selection. Only 5.3% of the children in our sample switch to a different school between spring of kindergarten and spring of first grade, most likely because almost all children in our sample attend schools with attached kindergartens. Thus, the vast majority of sorting into schools is already done by kindergarten and controlling for the kindergarten test score should adequately account for sorting of children into schools.

Section 7.1 provides empirical evidence that our strategy adequately deals with selection. First, our results are robust to both the exclusion and addition of important covariates like parental education and employment or other lagged test scores, all of which are likely to be related to this kind of selection. This suggests that selection on unobservables is unlikely. Second, our specification passes a falsification test using the kindergarten test score as the dependent variable.

5. Data

We estimate the model using data from the Early Childhood Longitudinal Study class of 1998–1999 (ECLS-K). We focus on reading and math test scores as our dependent variables, since these are commonly analyzed measures of achievement and can be compared over time. An advantage of the ECLS-K over commonly used administrative data sources (e.g. Hastings and Weinstein (2007); Jackson (2013); Chetty et al. (2014)) is the richer set of controls to separate the effect of school quality from past school histories, ability and family background. In addition to controlling for lagged test scores as discussed above, we condition on child characteristics (gender, race, birth weight, height, weight, BMI, disability), parents’ characteristics (socio-economic status, presence of father in the household, parents’ education and employment status, receipt of WIC benefits) and indicators for census region and urban location. We also control for variables related to the schooling history of the child, such as whether the child changed schools, age at entry into kindergarten, age at the time of the assessment and days elapsed since the first assessment to control for differential timing of the tests. See Appendix 1 for summary statistics of the variables we include as well as the time when they were measured.

6. Choice of proxies

We estimate the proxy variable model using four proxies each for teaching and resource quality. Including additional proxies does not affect the results substantively. We group the proxy variables based on the aspect of school quality one would expect them to be related to. That is, we use measures of teachers’ education and training as proxies for teaching quality, while we use the availability of staff, computers and other physical resources as proxies for resource quality. We follow the literature in restricting the proxies to be related to only one factor.

We use four teaching quality proxies related to teachers’ education and training aggregated at the school level to capture the quality of teaching in the schools. These include the average of all teachers’ years of schooling, the average number of college courses in reading/math taken by teachers, the proportion of teachers in a school who have elementary education certification and the proportion that have advanced professional certifications. Using averages across the teachers who were interviewed in a school allows us to get a measure of the latent teaching quality at the school level, which is determined by school level factors (which, in addition to the combination of all teachers, also include factors such as the principal) rather than individual teachers only. In addition, averaging these characteristics over multiple individual responses likely reduces measurement error due to survey errors. The teaching quality proxies we use have been widely analyzed in the school quality literature, though often with mixed results.11

Our second dimension of quality is a resources component, which captures the factors that schools with more resources can provide. We use average first grade class size, instructional computers per student, administrative staff per student, and student-specialized staff ratio as proxies for resource quality. Specialized staff include library/media, speech, reading, math, and foreign language specialists and nurses, which are resources that better off schools can provide more of. The measures we use as proxies for resource quality have also been commonly used in the literature with mixed results. Descriptive statistics of all proxies are shown in Table 1. The average years of teachers’ schooling in our sample is 15.8 while the average first grade class size is 21 students.

9 The controls in the two models are slightly different, because some controls matter for reading, but not for math. Using the same controls for both models does not change the results substantively, as can be seen in Appendix 2.

10 It is important to distinguish the teaching quality dimension we use and teacher quality measured by teacher fixed effects (e.g. Rockoff (2004)). Teacher fixed effects measure teacher-specific attributes that matter for child learning. Teaching quality here captures a component of teaching quality that is tied to the school instead of to specific teachers.

11 Existing estimates of the relationship between teacher’s schooling and student achievement are varied (Hanushek (2003); Past studies (Goldhaber and Brewer (2000); Darling-Hammond et al. (2001); Jepsen and Rivkin (2002)) are also mixed on the effect of teacher certifications on student achievement (Hanushek and Rivkin (2006)).

12 Several studies such as Angrist and Lavy (1999) and Chetty et al. (2011) find significant, positive effects of smaller class sizes. Dieterle (2015) documents that class size reductions can also have negative consequences, but contrary to him, we consider cross-sectional and not longitudinal differences in class size.

13 The increase in the number of non-instructional staff has been higher in schools in richer communities, while poorer communities have suffered cutbacks in such staff during budget crises (Tyack, 1992).
The model is robust to the inclusion and exclusion of specific proxies, i.e. the results are not driven by any individual proxy. Section 7.1 presents evidence against misspecification. The proxy correlations in Table 2 show that the proxies measure distinct factors that are correlated with each other. The low correlations suggest that school quality is hard to measure. Unlike factor models, our GMM model does not depend on the strength of the correlation between the proxies, so this does not affect the consistency of our estimates. The correlations are different across different pairs of proxies, which shows that classical measurement error is not a valid assumption, so IV estimates would be inconsistent. Therefore, we need a model like ours that allows for a more general form of measurement error.

7. Results

This section discusses the results of our model and provides empirical evidence of the validity of our assumptions and robustness checks in Section 7.1. We show that using a proxy variable model to estimate the effects of school quality affects the impact estimates substantively and provide illustrations to aid the interpretation of coefficients in Section 7.2. We discuss policy implications in Section 7.3.

Table 3 presents GMM estimates of the impact of two dimensions of school quality – teaching and resource quality – on reading and math test scores for students in first grade. First, we discuss the coefficients on the proxies in panel B of Table 3. The coefficients on all teaching quality and resource quality proxies are significant and have signs in the expected direction. Schools with teachers with higher average schooling and teachers who have taken more college level reading and math coursework have higher teaching quality. Larger class sizes are associated with lower resource quality, while higher ratios of instructional computers to students are associated with higher resource quality. The fact that all proxy coefficients have signs in the expected direction underlines that we are identifying factors that behave the way school quality would.

While some of the proxies may be direct inputs into student achievement, we cautiously interpret the proxy coefficients as associations. Since proxies were normalized to have a variance of one, the coefficients on the proxies represent the correlation between each proxy and the latent quality dimension, and the square of the coefficient is the fraction of the variance of the proxy that is explained by $Q$. The closer $\delta_{ij}$ is to one in absolute value, the less noise the proxy contains. Teachers’ schooling has the highest correlation with teaching quality, while administrative staff per student and class size are most strongly correlated with resource quality.

Turning to the impacts of school quality, we find that both teaching and resource quality are related to reading scores in first grade but do not find a significant effect on student achievement in math. All else equal, an increase in teaching quality by one SD is associated with 0.03 higher first grade reading test scores. The effect on first grade math scores is lower at 0.014 SD and not statistically significant, but still within sampling variation of the reading coefficient. An increase in resource quality by one SD is associated with a 0.04 SD improvement in reading and no discernible improvement in math – the point estimate is close to zero, negative and statistically insignificant. Our finding of significant effects on reading but no effects on math may reflect that reading achievement is the focus of early elementary education. For example, teachers in the ECLS-K report spending 90 min or more every day on reading instruction compared to only 30–60 min twice a week on math instruction (Croninger et al. (2007)). The estimated correlation between the two quality dimensions is negative. Such inverse relationships between school measures have previously been reported and may arise from compensating differentials or budget constraints (e.g. Heckman et al. (1996)).

7.1. Testing the model

The model specification and the identification of parameters rely on our controls being sufficient for assumption A1 to hold, the measurement system being well specified (assumptions A2 and A3), and the proxy variable model we propose working better than simpler models. This section provides formal tests and evidence that supports our model and specification choices.

Our specification of the observable part of Eq. (2) relies on a series of standard model specification tests. First, results in Appendix 2 show that our estimates are robust to changes in the covariates that one might

---

**Table 1**

Descriptive statistics of proxies.

<table>
<thead>
<tr>
<th>Teaching quality proxies</th>
<th>Resource quality proxies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Advanced professional certification</td>
<td>8.05%</td>
</tr>
<tr>
<td>Elementary education certification</td>
<td>92.09%</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>15.82</td>
</tr>
<tr>
<td>College courses in math</td>
<td>2.79</td>
</tr>
<tr>
<td>College courses in reading</td>
<td>3.63</td>
</tr>
<tr>
<td>N</td>
<td>10,059</td>
</tr>
</tbody>
</table>

Notes: Authors’ calculation using ECLS-K. Descriptive statistics for the proxies are provided for the estimation sample used for math test scores as it is the larger sample. The estimates for the sample used for reading test scores are similar and included in Appendix 1.

**Table 2**

Correlations between school quality proxies.

<table>
<thead>
<tr>
<th>Teaching quality proxies</th>
<th>Resource quality proxies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced professional certification</td>
<td>0.012</td>
</tr>
<tr>
<td>Elementary education certification</td>
<td>0.249***</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>0.190***</td>
</tr>
<tr>
<td>College courses in math</td>
<td>0.103***</td>
</tr>
<tr>
<td>College courses in reading</td>
<td>0.192***</td>
</tr>
</tbody>
</table>

Notes: Authors’ calculation using ECLS-K. The correlations presented are for the sample used in estimation for math test scores as that is the larger sample. The correlations for the reading test sample are very similar, since the math and reading test samples are so similar. †: significant at 5%; ‡: significant at 1%; ***: significant at 0.1%.

---

14 All teaching proxies are positively correlated with each other. Some pairwise resource proxy comparisons are negative, but only those where it is plausible that one of the proxies is negatively related to school quality.

15 Estimates from one-step GMM are almost identical to our optimally weighted GMM estimates, so small sample bias (Altonji and Segal (1996)) does not seem to be an issue.
expect to be related to selection. Once the model includes the lagged test score in the same subject, the coefficients stabilize quickly and are insensitive both to adding important covariates (such as the lagged test score in the other subject) and to removing important covariates (such as parents’ education and employment status). If the model were not well specified, one would expect the results to be sensitive to the inclusion of variables related to selection. This suggests that our conditioning set likely solves the selection problem and that the observable part of Eq. (2) is correctly specified.

We also run a falsification test by using the baseline score from fall kindergarten as the dependent variable in our model. This entering kindergarten test score captures pre-school achievement, so it should not have been meaningfully impacted by school quality. Therefore, a significant relationship between it and school quality would suggest that the controls are not sufficient to account for selection. However, the results in Appendix 3 are not significant, which further suggests that our conditioning set is sufficient to capture selection into schools.16

The results above provide evidence that our estimates would be unbiased if school quality were observable, but our model also rests on the specification of the proxy variable model, i.e. assumptions A2 and A3, and that Eq. (2) is linear in $Q_t$. To evaluate these assumptions, we use the fact that the proxy variable model we propose implies a specific relationship between the results we obtain and the results that would be obtained when ignoring the measurement error problem and simply including the proxies in OLS. Black and Smith (2006) provide the following predictions if the latent quality approach is correct: first, proxies with a higher factor loading, $\beta_{it}$ (i.e. sending a stronger signal) should have estimated OLS coefficients with lower p-values when entered one by one and, second, estimates should become more attenuated when proxies are entered pair-wise in OLS.17 While the proxies with the highest factor loadings for each quality dimension in our GMM estimation do not always have the lowest p-value in the OLS regressions, the pair-wise regression results attenuate as expected. The coefficients on 19 of the 24 proxies for reading and 15 of the 24 proxies for math are attenuated in pair-wise OLS regressions relative to the coefficients from OLS regressions of test scores on the individual proxies.18

We also perform a formal specification test in the spirit of Hausman (1978). The parameter estimates of our measurement error model can be used to calculate the probability limit of the implied OLS coefficients when the proxies are used in OLS. Our test statistic is the difference between these implied OLS coefficients and the OLS coefficients one obtains when actually regressing $\theta_t$ on $X_t$, $\theta_{t-1}$, and one proxy per latent dimension. Under the null hypothesis that our model is correctly specified (i.e. A1–A3 hold and Eq. (2) is linear in $Q_t$), this difference should only differ from zero by sampling variation. However, one would not expect the difference to be zero if the model is misspecified, since the relationship between implied and actual coefficients depends on the number of dimensions, the linearity assumption in Eq. (2), as well as assumptions A1–A3. Standard errors are calculated using a cluster-robust pairs-bootstrap (Cameron et al. (2008)). See Appendix 4 for further details on the test.

Table 4 shows that this test does not reject the null hypothesis for any of the proxies for reading and for seven out of eight proxies for math. For math, we reject the null hypothesis for the average first grade class size proxy. However, finding that the difference between the actual and implied coefficients is statistically significant for one out of 16 coefficients is within sampling variation.

While this test has the advantage of jointly testing all assumptions of the model, its power against specific alternatives is unclear. We conducted simulations in which either A1 does not hold (because the true model includes squared or interaction terms in $Q_t$) or A3 is violated (because errors in the proxy equations are correlated with each other or with $\epsilon_t$, the error in the outcome equation). We find that the test is sensitive to both types of violations of assumption A3 but may not have power to detect violations of A1 where $Q_t$ enters Eq. (2) non-linearly.19 Further analyses, such as estimating a factor model with an

---

16 This test differs from our main specification as we do not have lagged test scores here. If both lagged and current scores are positively correlated with school quality, this creates an upward bias making it harder to pass the test.

17 These predictions do not necessarily hold when there are two latent dimensions, since the bias from including one mismeasured variable affects the coefficient on the other mismeasured variable and may undo these predictions. However, since a tendency towards attenuation should persist and affect the coefficients with lower factor loadings more, this exercise can be interpreted as informal evidence for our specification.

18 The four proxies per dimension can be arranged in 6 pair-wise combinations. Each pair-wise combination yields 2 coefficients and there are two quality dimensions for a total of 24 coefficients per test score.

19 Simulation results when A3 is violated are included in Appendix 4.
additional dimension, which often captures a non-linear impact, do not yield any evidence of non-linear relationship. It is unlikely that we would fail to detect a strong non-linear impact, so even if there are small non-linearities that violate assumption A1, we do not expect them to affect our conclusions substantively.

Regarding assumption A3, the simulations show that the test has power against violations of this assumption, which is much harder to assess substantively. To further assess the validity of A3, we compute lower bounds for the impact of teaching and resource quality using the estimator in Lubotsky and Wittenberg (2006) which provides a lower bound if A3 is violated (Black and Smith (2006)). Unfortunately, the lower bounds for resource quality are close to zero and, therefore, not informative about whether the effect of resource quality is zero or positive. However, the lower bounds for the effect of teaching quality exceed our estimates slightly and, therefore, suggest that if our positive and significant estimate of the impact of teaching quality on student achievement is biased, it underestimates the true impact.

In summary, neither the informal evidence nor the formal tests discussed in this section contradict the assumptions our latent quality approach estimation relies on. Thus, we take the overall evidence to suggest that our model is well specified.

7.2. Interpretation of effects

Having discussed several pieces of evidence that support the proxy variable model in Section 7.1, we turn to examining whether the results are substantively affected by using this proxy model rather than linear regressions, and put the effects we find into context. We illustrate the differences in latent school quality in terms of the proxies used, and relate the impact estimates found to the size of the black-white test score gap. These exercises are meant to provide context and do not imply causality.

We find that ignoring the proxy variable formulation leads to considerably biased estimates of the impact of school quality. Following Black and Smith (2006), Table 5 shows that the OLS estimates that directly relate the proxies to reading test scores are attenuated. All estimates bear the expected sign but they are insignificant in seven out of eight regressions. For example, directly using teachers’ schooling as a measure of teaching quality in OLS yields an estimate that is statistically insignificant and 50% smaller than the impact of teaching quality in our proxy variable model. Averaging across the teaching quality proxies, we find that the coefficients of teaching quality proxies in OLS are attenuated by 44% on average. The analogous attenuation bias when using a resource quality proxy in OLS is 58%. As we discuss below, these biases have implications for cost benefit analyses, which underscores the importance of taking the measurement error in the proxies into account.

A quality score can be constructed for each school with our estimates, but it would require that we assume that our controls (X, X−1) are orthogonal to latent quality (Qt). We take a different approach in order to avoid this assumption. We analyze how the teaching and resource proxies vary across schools with different latent quality holding student composition (as measured by X, X−1) fixed. This can be seen as a thought experiment where we analyze the difference in proxies as the latent quality changes among schools with the same student composition in terms of our controls. Note that Q′ enters all four proxy equations, so increasing it by one unit increases the expected value of all associated proxies. Since we estimate the effect of teaching and resource quality on achievement, we show the illustration for both dimensions of school quality. This exercise also provides suggestive evidence of factors that influence these dimensions of school quality and ways to detect good schools, both of which are important for policy.

Table 6 shows that, conditional on having the same student composition, schools with higher teaching quality tend to have higher levels of all four teaching proxies. For example, a school with a one standard deviation higher teaching quality tends to have four percent more teachers with advanced professional certification and teachers with almost one additional year of schooling, on average. The differences in the other proxies can be similarly determined from the table below where the mean of each variable is provided as a reference. Schools that are one standard deviation higher in terms of resource quality typically have 1.6 fewer students per class in first grade.

Having shown that estimation of the proxy variable model yields significant positive impacts of school quality on reading test scores, we benchmark these effects to the unadjusted black-white test score gaps reported by Fryer and Levitt (2004) using the ECLS-K data. While we do not find statistically or economically significant effects of school quality on math achievement, we find that an increase of one standard deviation each in teaching quality and resource quality between spring of kindergarten and spring of first grade is associated with an improvement of roughly 0.07 SD in reading test scores. Fryer and Levitt find that the unadjusted black-white reading test score gap is 0.4 SD in the fall of kindergarten and 0.529 SD by the spring of first grade. The effect of

---

21 Since we only find significant quality impacts on reading and not for math, we document the attenuation bias from OLS estimation for reading test scores only.

22 For the sake of brevity, we use reading scores only. The school quality score constructed for reading and math differ slightly because the samples are slightly different and the teaching proxy of average college courses in reading/math differs for the two subjects, but are highly correlated (0.95 for teaching quality and 0.99 for resource quality).
teaching and resource quality we find is important, as the effect of increasing both teaching and resource quality by one SD is more than half of the 0.129 SD that white children gain relative to black children between fall of kindergarten and spring of first grade. While the unadjusted test score gaps may capture several differences across the children, Fryer and Levitt also argue that blacks attending schools of worse quality than whites is already an important explanation for this worsening trajectory of black children.

The effects we have found are likely to underestimate the importance of school quality. We have estimated the impact of school quality in first grade alone, i.e. over a period of just one year between spring of kindergarten and spring of first grade. The effect of improved teaching and resource quality for a longer period of time would presumably be larger. The effects in first grade can also be amplified in the longer run if skills beget skills (Cunha et al. (2010)).

7.3. Policy implications

Test score improvements are valued for their impact on later life outcomes including higher earnings. We provide a rough, back-of-the-envelope calculation to indicate the implied earnings returns of improved school quality through its effect on reading test scores. The purpose of this exercise is not to provide a convincing estimate of earnings returns to improved school quality. Rather, the goal is to put into context the importance of the effect size we find. Using the estimates of Chetty et al. (2011) on the relationship between kindergarten test scores and adult earnings suggests that an improvement of first grade reading scores by 0.07 SD adds an expected $4773 in lifetime earnings for each child.23 The average first grade class contains 20 students. Therefore, an increase of one SD each in teaching and resource quality at the school level implies earnings gains of approximately $106,377 per classroom in first grade alone. The impact of school quality over more grades would presumably be larger. If we had ignored the proxy variable formulation of the model and used OLS, the estimated earnings gains due to improvement of school quality in first grade would be almost halved, at $58,043.24 In line with Chetty et al. (2011), our findings suggest an important relationship between school quality and first grade test scores with substantial impact on future earnings. While we do not attempt to estimate the production function of school quality, our results suggest that analysis of this production function is important.

8. Conclusion

Using a proxy variable model, we find a significant and small but substantively important association between school quality and student achievement at the individual level. While we do not find statistically significant effects of school quality on math, the estimated effect of school quality on reading achievement is statistically significant and important. We have subjected our model to a number of specification tests and conclude that it performs better than common regression models that do not take the measurement error in the available school quality measures into account.

Our results provide an important step in establishing a connection between school characteristics, school quality, and student outcomes in micro data and thereby uncovering the channels through which better schools lead to better outcomes later in life. Our results show that school quality influences test scores and therefore needs to be accounted for in studies of skill production to avoid bias. We also provide evidence that commonly used indicators of school quality are noisy proxies and that the measurement error leads to bias if it is not modeled appropriately. This underscores the importance of using proxy variable methods to assess the impact of school quality on achievement, and may help to explain why findings from aggregate and individual level studies tend to diverge. While our study has focused on the effect of school quality in first grade on reading and math achievement, the same approach could be used to explore whether the impact is different for other age groups, and whether there are other components of school quality besides teaching and resource quality.

Acknowledgements

We would like to thank Guadalupe Bedoya, Dan Black, James Heckman, Robert Lalonde, Ofer Malamud, Steve Rivkin, Peter Saveliev and seminar participants at the University of Chicago's Workshop for Human Potential and the Harris School PhD. workshop for providing helpful comments and Deborah Novakova for editing assistance. The suggestions from the editor, Helena Skyt Nielsen and three anonymous referees have greatly improved this paper. Any remaining errors are the sole responsibility of the authors. Mittag acknowledges support by grant P402/12/G130 from the Czech Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.labecono.2016.01.005.

References


Table 6

Change in proxies associated with a one standard deviation change in latent quality (\(\Delta q\)).

<table>
<thead>
<tr>
<th>Teaching quality proxies:</th>
<th>Mean</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced professional certification</td>
<td>0.082</td>
<td>0.042</td>
</tr>
<tr>
<td>Elementary education certification</td>
<td>0.928</td>
<td>0.031</td>
</tr>
<tr>
<td>Years of schooling</td>
<td>15.8</td>
<td>0.927</td>
</tr>
<tr>
<td>College courses in reading</td>
<td>3.653</td>
<td>0.489</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource quality proxies:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average first grade class size</td>
<td>20.72</td>
<td>−1.629</td>
</tr>
<tr>
<td>Instructional computers per student</td>
<td>0.183</td>
<td>0.023</td>
</tr>
<tr>
<td>Administrative staff per student</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>Student-specialized staff ratio</td>
<td>290.31</td>
<td>−69.538</td>
</tr>
</tbody>
</table>

Notes: Authors’ calculations using ECLS-K. Change in school characteristics is conditional on the controls we use in our model and does not represent a causal relationship. The teacher certification proxies represent proportions of teachers.

23 Chetty et al. (2011) report that a one standard deviation increase in test scores is associated with a $2804.16 (18%) increase in earnings. We calculate lifetime earnings gains using forty years of earnings and a 3% discount rate.
24 We use the average attenuation bias from using individual teaching and individual resource proxies as measures of school quality to calculate the implied OLS impact estimates and implied earnings returns.