Swedish Institute for Social Research (SOFI)

Stockholm University

WORKING PAPER 1/2016

WHAT DO BOOKS IN THE HOME PROXY FOR?
A CAUTIONARY TALE

by

Per Engzell

What Do Books in the Home Proxy For? A Cautionary Tale

Per Engzell

Swedish Institute for Social Research (SOFI), Stockholm University
per.engzell@sofi.su.se

February, 2016

Abstract

A large body of work in the social sciences relies on proxy variables to capture the influence of an unobserved regressor. Assuming that measurement error is well approximated by a classical model implying bias toward the null, proxies that explain a larger amount of variance in the regression are routinely preferred. I show how this reasoning can mislead, examining a widely used predictor of student achievement: the self-reported number of books at home. Underreporting by low achievers and endogeneity of parental inputs both contribute an upward bias, large enough to overturn the classical attenuation result and lead to spurious inferences. The findings serve as a caution against overreliance on standard assumptions and cast doubt on predictive power as a criterion for proxy selection.

Keywords: *education, equality of opportunity, socioeconomic status, proxy variables, differential measurement error*; JEL: *C81, I21, I24, J62*

I. Introduction

In the years leading up to 1915, Charles Elmer Holley, a doctoral candidate at the University of Illinois, surveyed students and their parents in high schools throughout the state. In his thesis submitted that year and issued as a *Yearbook of the National Society for the Study of Education* the following, he wrote:

> If a person wished to forecast, from a single objective measure, the probable educational opportunities which the children of a home have, the best measure would be the number of books in the home. (Holley 1916, p. 100)

His conclusion was based on cross-tabulations and bivariate correlations involving offspring's years of schooling and various family characteristics. Holley granted that his data were likely not without errors of observation, but believed that the consequence would be "nearly that of pure chance, though this may be proved otherwise if carefully investigated" (p. 17).

Measurement of parental background is crucial in research on educational production, skill formation, and socioeconomic differences in achievement (Björklund and Salvanes 2011, Hanushek and Woessmann 2011, Heckman and Mosso 2014). Virtually without exception, methodological literature on proxy variables departs from some version of the 'classical' model where error is

treated as Gaussian noise, or more recently, the weaker assumption that it is nondifferential: unrelated to regression residuals. It is well known that such error will lead to a bias toward the null, in the classical case as a simple function of the ratio of noise to total variance (Griliches 1986). Guided by this heuristic, researchers often deliberately seek proxies that account for the largest amount of variance in the dependent variable, with the implication that they more closely track the attributes proxied for, are more reliably reported, or both. Mostly this practice is implicit, but formal arguments are found in Leamer (1983) and Lubotsky and Wittenberg (2006).

In this context, the number of books in the home (henceforth, NBH) has gained considerable popularity,[1] especially in surveys involving school children who may have difficulties reporting their parents' education or income. Hanushek and Woessmann (2011, p. 117) describe NBH as "a powerful proxy for the educational, social, and economic background of the students' families". It is one of few socioeconomic status (SES) indicators consistently available across international student assessments, and has been a staple of these surveys since their inception (Thorndike 1973). It also appears in U.S. Department of Education studies including the National Assessment of Educational Progress (NAEP). Beyond its use as a standalone proxy, it figures in widely used factor-based measures such as the Index of Economic, Social and Cultural Status (ESCS) in OECD's Programme for International Student Assessment (PISA), or the similar scales in the studies of the International Association for the Evaluation of Educational Achievement (IEA).

Interestingly, the perception that children are able to report reliably on NBH stems not so much from direct evidence as from the strong associations noted already by Holley. In this vein, Hanushek and Woessmann (2011, p. 117) recommend NBH as a catch-all for socioeconomic background "not only because cross-country comparability and data coverage are superior . . . but also because books at home are the single most important predictor of student performance in most countries". A 100-page methodological monograph on the subject recently issued by the IEA and the Educational Testing Service (ETS) similarly urged survey organizers to include those measures "that show the highest association with achievement in terms of explained variance",

[1] As of this article's writing, a search for "number of books [at/in the] home" returned over 3,500 results in Google's *Scholar* database, two thirds of which were penned in the last decade. These writings span the social sciences including economics, psychology, sociology, and educational research. The claim that NBH is the "single most important" predictor of achievement is a reoccurring one (Algan, Cahuc, and Shleifer 2011, Ammermueller and Pischke 2009, Hanushek and Woessmann 2011, Peterson and Woessmann 2007, Schütz, Ursprung, and Woessmann 2008). The strong associations have also been reported in popular media, influencing public discourse (e.g., *The New York Times*, 2011, 2015a,b). Evidence on cross-country differences, peer effects, or the impact of tracking drawing on self-report NBH is cited in handbook chapters by Betts (2011), Epple and Romano (2011), and Hanushek and Woessmann (2011). A selective bibliography includes Algan, Cahuc, and Shleifer (2013), Ammermueller (2007, 2013), Ammermueller and Pischke (2009), Brunello and Checchi (2007), Brunello, Weber, and Weiss (2015), De Witte and Kortelainen (2013), Ferreira and Gignoux (2014), Freeman, Machin, and Viarengo (2010), Freeman and Viarengo (2014), Fuchs and Woessmann (2007), Jürges, Schneider, and Büchel (2005), Martins and Veiga (2010), Ohinata and van Ours (2013), Peterson and Woessmann (2007), Raitano and Vona (2013), Schneeweis and Winter-Ebmer (2007), Schütz et al. (2008), Waldinger (2006), and Woessmann (2003).

identifying NBH as "the strongest predictor of achievement . . . across the different studies and subject areas investigated" (Brese and Mirazchiyski 2013, pp. 98-99).

A related example of reliance on classical assumptions is Ammermueller and Pischke (2009) who, in their important work on classroom peer effects, draw on separate reports by students and parents in the Progress in International Reading Literacy Study (PIRLS) to correct for attenuation in NBH. Employing a standard instrumental variable (IV) solution, they see their estimates triple in size. Reflecting much of the literature, they discuss mean reversion (negative correlation between the error and true values) as a potential threat to identification, but do not take issue with the assumption of nondifferentiality underlying nearly all methodological work. Their attention to measurement error is stressed as a key contribution, a sentiment echoed in a literature review by Epple and Romano (2011).

Revisiting the PIRLS data, I develop a method to decompose the bias in self-reported books allowing for a rich error structure. I find that much of the measure's association with achievement appears driven by differential misreporting, with low achievers tending to underestimate the number. An additional source of endogeneity is reciprocal causation, as parents acquire more books for students who are better readers. The resulting upward bias is severe enough to eclipse attenuation from random noise, and lead to a net bias of unpredictable direction and magnitude. Implications are similar to those of any standard endogeneity problem, but with the exception that IV is generally not a solution: the critical error is both endogenous and mean reverting, biasing IV and OLS estimates alike (Kane, Rouse, and Staiger 1999).

These findings illustrate the need for researchers to check and justify, on a casewise basis, the conditional independence assumptions used in dealing with proxy variables and measurement error. They also raise concerns about a recent proposal by Lubotsky and Wittenberg (2006) on how to extract structural parameters when several proxies for an unobserved regressor are available. The Lubotsky–Wittenberg procedure generalizes the practice of maximizing variance explained to the case of multiple indicators. Such an approach is going to prove useful if there are strong reasons to assume nondifferentiality in each measure. Usually this assumption is invoked as a matter of routine, however, suggesting that the 'optimal' weighting might capitalize on chance violations of it.

One reason why researchers have often been reluctant to assess differential error might be a belief that the assumption is untestable. I show that this is misguided: much can be learned from information about the predominant direction of error, or differences along auxiliary variables that are unrelated to the regressor of interest. In the setting studied here, gender differences in reported books provide such an example because gender is strongly correlated with reading achievement at young ages but not with student background. Applying this method corroborates the main results, but also suggests that student reports of parental education is subject to the opposite error – low

achievers tend to exaggerate this variable, biasing associations further downward – while use of parental occupation largely avoids these problems.

In what follows, Section II provides further background on NBH and related proxies. Section III details the formal framework guiding the analysis. Section IV introduces the data and inspects NBH through a series of simple checks which, if followed, would have prevented its misuse. Section V describes and applies a method to decompose the bias from student-reported data allowing for arbitrary error, as well as error of a known structure in the validation reports by parents (motivated below). Section VI concludes by drawing lessons for the study of socioeconomic achievement gaps and research using survey-reported proxy variables more generally.

II. Background and Previous Literature

This section describes the background and uses of NBH; the general problem of differential measurement error is discussed in greater detail in the next.

Associations between family background and student achievement are a matter of concern for policymakers, scholars, and the general public. A widely espoused ideal holds that a person's chances to get ahead should depend not on accidents of birth but rather on talent and hard work. As early achievement is a strong determinant of later economic success, compensating for differences that are present at this stage, thereby 'leveling the playing field', is an important objective. Recent international learning assessments have spurred research in this field by increasing the amount of available data.[2] This research relies on proxies such as parents' education, income, or – as here – home library to capture the various parental inputs that are thought to matter.

Because crude associations partly reflect mechanisms that are not feasible or desirable to eliminate (inherited differences in ability, for example), a common strategy is to focus on *variation* in associations to draw conclusions of relevance to policymakers. As Hanushek and Woessmann (2011, p. 123) state: "lacking obvious reasons to assume that natural transmission differs across countries, cross-country comparisons can be interpreted in terms of differences in the extent to which societies achieve more or less equal educational opportunities". In addition to the underlying processes being invariant, we also have to assume that the relationship between the proxy and its proxand(s) is stable, and that measurement quality does not differ markedly across contexts.

Most discussions of measurement error rely, implicitly or explicitly, on the classical assumptions that the error is mean zero, normally distributed, and uncorrelated with true values as well as

[2] The Programme for International Student Assessment (PISA) is carried out every three years since 2000 by the Organisation for Economic Co-operation and Development (OECD). Its results have been widely publicized and prompted policy responses in several participant countries. The Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) are carried out by the International Association for the Evaluation of Educational Achievement (IEA), currently every four (TIMSS) or five (PIRLS) years. While IEA's history of assessment extends as far back as the late 1950s, TIMSS and PIRLS found their current form in 1995 and 2001.

all other variables in the model. These assumptions greatly simplify estimation. If signal-to-noise ratios are constant, for example, relative comparison of coefficient sizes will be unaffected by error (Jerrim and Micklewright 2014). This applies to descriptive evidence across countries, student ages, or outcome domains (Woessmann 2003, Schütz, Ursprung, and Woessmann 2008, Martins and Veiga 2010), and to estimated intervention effects (Waldinger 2006, Brunello and Checchi 2007, Ammermueller 2013). The classical model or nondifferential error restriction more generally is also crucial to conventional methods of bias correction (Bound, Brown, and Mathiowetz 2001), and to the identification of social interaction effects (Epple and Romano 2011).

The quality of social background variables collected from students has been an active research area at least since the 1970s (e.g., Mason et al. 1976). This literature, focusing mostly on parents' education and occupation, concluded that while education is often reported with considerable error, students as young as ten are able to report their parents' occupation with some accuracy (Looker 1989). More recent research substantiates these findings (Jerrim and Micklewright 2014, Engzell and Jonsson 2015). A separate research strand has explored measures based on household items, with somewhat mixed results (Traynor and Raykov 2013).

While many studies cite explanatory power or favorable response rates in support of NBH, direct evidence on its validity or reliability is sparse. Schütz et al. (2008) regressed a banded measure of annual household income on NBH using parent-reported data from 6 countries in PIRLS 2001. They interpret the absence of significant country interactions in this regression as "strong evidence [of] the validity of cross-country comparisons where the books-at-home variable proxies for family background" (pp. 287-288). The power of this test is questionable since income is itself volatile and typically reported with much error (Micklewright and Schnepf 2010). But more fundamentally, because data were sourced from parents, the evidence does not speak to the quality of student reports, which is what most studies (including Schütz et al.) ultimately have relied on.

A handful of studies indicate that parent–student agreement on NBH is low, but fail to reconcile this with the sizeable outcome associations usually estimated (Jürges and Schneider 2007, Rutkowski and Rutkowski 2010, Jerrim and Micklewright 2014). Jürges and Schneider (2007, p. 421) argue that if either source was less reliable it "should have a smaller correlation with the dependent variable" which they note is not the case. The most careful validation to date is by Jerrim and Micklewright (2014) who show that regression estimates using NBH are quite volatile depending on the source of reporting and advise caution. They discuss differential error, but conclude (erroneously) that if "more able children provide better reports . . . a [further] downward impact on the estimate" will result (p. 780, cf. Kreuter et al. 2010).

The research referenced here obviously differs in assumptions from a related literature that has drawn on longitudinal data and explicitly modeled the endogeneity of parental inputs such as books (e.g., Cunha, Heckman, and Schennach 2010). One reason why exogeneity has seemed

plausible in the cross-sectional case is that the item used typically refers to the total amount of books in the home,[3] unlike longitudinal studies that tend to track children's books specifically. So, for example, Ammermueller (2007, p. 247) argues that, along with parents' education and country of origin, "books at home ... are unlikely to change over time and may serve as a good proxy for prior inputs". An important lesson in the following is that even if books were exogenous, student self-reports would still be endogenous because response errors are not random.

III. Formal Framework

The problem we are concerned with is to estimate the parameters in a regression of reading test scores on the number of books in the home. The issue of how NBH relates as a proxy to the underlying parent characteristics, while important, is beyond the scope of this paper so I will take for granted that in the absence of differential error, NBH would track the concept of student background we are after. Assuming $\mathrm{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ we write the target regression:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \perp \boldsymbol{X} \tag{1}$$

where $\boldsymbol{y}$ is a test score and $\boldsymbol{X}$ a set of predictors including NBH, $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are column vectors of dimension n, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are of dimension $n \times k$ and $k \times 1$, and one of the k columns (rows) is reserved for the intercept. As the books question is usually categorical ("0–10 books", "11–25 books", etc), to abstract from errors due to truncation or discretization we assume that the categories and not the underlying continuous variable are the target. Empirically, $\mathrm{E}(y_i|\boldsymbol{x}_i)$ tends to be roughly linear in categories so NBH is often used as if it were a continuous variable.[4] For parsimony this practice will be followed here, but the framework applies equally to a categorical (dummy variable) specification, with or without additional covariates.

In keeping with the literature, $\boldsymbol{X}$ in the regression above is assumed exogenous. This does not mean that the expectation function $\boldsymbol{\beta}$ has a direct causal interpretation, only that $\boldsymbol{X}$ does not change as a consequence of $\boldsymbol{\varepsilon}$.[5] While this is reasonable with proxies such as parental education or occupation, it is more problematic for NBH. It is easy to imagine that book consumption is a function not only of predetermined parental characteristics but also of the student's interest and aptitude in reading. We must therefore think of the variable as the number of books before the

[3] An exception is Fryer and Levitt (2004) who note the considerable explanatory power of children's books in the Early Childhood Longitudinal Study and conclude that the variable "seems to serve as a useful proxy for capturing the conduciveness of the home environment to academic success" (p. 452).

[4] For example, Schütz et al. (2008), Ammermueller and Pischke (2009). The problem of dicretized regressors is considered by Hyslop and Imbens (2001) and Manski and Tamer (2002), and that of the aggregation of several proxies by Black and Smith (2006) and Lubotsky and Wittenberg (2006).

[5] That is, we are not looking to estimate the change in reading proficiency that would result from endowing a home with n additional books. For Holley (1916, p. 63), NBH is "a rough index of the culture of the home", similar views are found in Schütz et al. (2008), Ammermueller and Pischke (2009), and Hanushek and Woessmann (2011).

student came of reading age, or (more pedantically) the expected number of books at the time of survey, given parents' permanent characteristics. In reality, we observe not $\boldsymbol{X}$ but a noisy measure of a potentially endogenous variable:

$$\boldsymbol{M} = \boldsymbol{X} + \boldsymbol{\xi} + \boldsymbol{\eta}$$

where $\boldsymbol{\xi}$ is a stochastic component of books and $\boldsymbol{\eta}$ is a response error reflecting the fact that the student may be misinformed, miscomprehend the question, or otherwise state the wrong answer. We write the total error as $\boldsymbol{U} = \boldsymbol{\xi} + \boldsymbol{\eta}$, and contrary to common practice, we will *not* assume that $\boldsymbol{U}$ is unrelated to either $\boldsymbol{X}$ or $\boldsymbol{\varepsilon}$. In this case, the ordinary least squares estimator equals (cf. Bound et al. 1994):

$$\begin{aligned}
\boldsymbol{b}_{OLS} &= \boldsymbol{\Sigma}_M^{-1}\boldsymbol{M}'\boldsymbol{y} \\
&= \boldsymbol{\Sigma}_M^{-1}\boldsymbol{M}'(\boldsymbol{M}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + \boldsymbol{\Sigma}_M^{-1}\big[\boldsymbol{M}'(-\boldsymbol{U}\boldsymbol{\beta}) + \boldsymbol{M}'\boldsymbol{\varepsilon}\big] \\
&= \boldsymbol{\beta} + \boldsymbol{\Sigma}_M^{-1}\big[\boldsymbol{M}'\boldsymbol{U}(-\boldsymbol{\beta}) + \boldsymbol{\xi}'\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\boldsymbol{\varepsilon}\big] \\
&= \boldsymbol{\beta} + \boldsymbol{\Sigma}_M^{-1}\boldsymbol{M}'\boldsymbol{U}(-\boldsymbol{\beta}) + \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{\xi}'\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\boldsymbol{\varepsilon})
\end{aligned} \tag{2}$$

Subtracting $\boldsymbol{\beta}$ from both sides and taking plims, the resulting bias becomes:

$$\begin{aligned}
&\operatorname{plim}\ \underbrace{\boldsymbol{b}_{OLS} - \boldsymbol{\beta}}_{\text{bias}} \\
= &\operatorname{plim}\ \underbrace{\boldsymbol{\Sigma}_M^{-1}\boldsymbol{M}'\boldsymbol{U}(-\boldsymbol{\beta})}_{\text{attenuation}} + \underbrace{\boldsymbol{\Sigma}_M^{-1}(\boldsymbol{\xi}'\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\boldsymbol{\varepsilon})}_{\text{endogeneity}}
\end{aligned} \tag{3}$$

The right-hand side contains two terms. The first involves a $k \times k$ matrix of stacked coefficient vectors that result from regressing each column of $\boldsymbol{U}$ on observed variables $\boldsymbol{M}$, multiplied with $-\boldsymbol{\beta}$. Bar unusual circumstances, the bias from this expression is toward the null for any mismeasured regressor and similar to that from classical errors in variables. Indeed, if there is only one regressor and its error is unrelated to true values, it simplifies to the classical attenuation bias: $-\beta\big[\sigma_u^2/(\sigma_x^2 + \sigma_u^2)\big]$. Mean reversion (negative correlation with true values) is subsumed in this term and, if present, will reduce attenuation.

The second expression involves two covariance terms $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$ that relate each of the errors to the equation disturbance. Conventional measurement error models simply assume that this expression is nil; whether this is reasonable or not is exactly what we want to find out. The suspected sign on $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$ is perhaps obvious. Some books will be brought into the house by the student, either directly or indirectly, and that number is likely higher for a student with a gift for reading. If so, $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$ is positive and contributes an upward bias.

What to expect of $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$ is less obvious, but a useful starting point is to assume that low achieving students report less accurately. The direction of bias is then determined by whether under- or

overreports dominate the error. There is previous evidence of overreporting or 'upgrading' in student reports of parental education (Kerckhoff, Mason, and Poss 1973). If related to low achievement, this means that $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$ is negative and will contribute a downward bias. But for NBH, we might as well expect the opposite: a student with little ability or interest in reading would seem likely to *under*estimate the number for the simple reason that s/he is unaware of most books in the home. If so, $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$ is positive and will contribute an upward bias like $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$.

In sum, error ridden regression estimates will be subject to bias toward the null due to the random component of error. Most discussions of measurement error stop there, and assume that such estimates will, in effect, be underestimates. But for NBH, this might be counteracted by the covariance terms $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$ which are plausibly positive: part of the actual variation in books will be endogenous, and part of the variation in reported books might be endogenous if low achievers underestimate the number. The direction and size of the bias is therefore indeterminate and cannot, in general, be inferred from estimated error rates.

IV. Books at Home in PIRLS

The following section introduces the PIRLS data and documents two important descriptive findings. First, using parents as a benchmark, underreporting of NBH is much more common than its opposite, and clearly associated with reading achievement. Second, relying on gender as a quasi-instrument for achievement reveals that boys report lower values. The same gender pattern holds to a lesser extent for parent reports about children's, but not non-children's books, and generalizes to the older PISA students.

IEA has collected data on books from students and parents in PIRLS every five years since 2001. The 2011 round was carried out on school-based, random samples of fourth-grade students (age 10) in near 50 countries. A parent questionnaire (the "Learning to Read Survey") was administered in 45 countries, but with poor response rates (below 60%) in 5 of them. I focus on the remaining 40, a list of which is provided in Table 2, all with parental response above 75%. In these countries, a total 222,425 students were assessed. Restricting the analyses to complete cases, where both the student and parent reported, yields a sample of 197,387.

As a first approximation it is useful to assume that parent reports are, if not correct, then at least much more accurate. Parents will, as adults, be better at the cognitive tasks involved in responding. They will also be better informed because they, not the student, have brought most of the books into the house and will have some attachment to them. Finally, parents answer the survey at home which should lead to more accurate answers about the home environment. This assumption will be subject to a sensitivity analysis below. I also draw on exogenous variation in achievement due to student gender as an alternative source of validation.

Table 1 shows the questions asked about NBH. While students are asked to estimate the total number of books, the parent questionnaire splits this item into "books" and "children's books".[6] The parent, but not the student, questionnaire also includes questions about parents' education, employment, and line of work. The same questions are used in IEA's other assessment, the Trends in International Mathematics and Science Study (TIMSS), where a parent questionnaire was first introduced in 2011. The third major assessment, OECD's PISA survey, asks students but not parents about books. While all these studies survey parents in some form, inconsistent coverage and varying response rates entail that for many purposes, student self-reports are the only viable source.

The concept of reading literacy in PIRLS is broad and includes comprehension as well as "the ability to reflect on what is read and to use it [to attain] individual and societal goals" (Mullis et al. 2009, p. 11). To assess a range of capabilities, a rotated booklet design is used where each student is tested on two out of a total ten text passages. Test scores are then imputed as posterior draws from estimated ability distributions following a Rasch model. I standardize these values to have mean=0, s.d.=1 within each country. Estimates are accounting for the uncertainty associated with plausible value imputation as well as for clustering on school classes. To economize on precision, survey weights are not applied (cf. Solon, Haider, and Wooldridge 2015) under the assumption that the qualitative results do not hinge on the representativeness of the sample; Stapleton and Kang (2016) investigate this issue with TIMSS data and suggest that the difference is minor.

A. Low Agreement between Students and Parents

I first revisit the low agreement found in previous studies on PIRLS data. Figure 1, left panel, plots Cohen's κ (kappa), a common measure of interreporter agreement, calculated separately for each country. The statistics are trailing well below the .40 threshold commonly taken to denote 'moderate' agreement (Landis and Koch 1977). As noted by Jerrim and Micklewright (2014) who estimate agreement on parental education and occupation from older PISA respondents, young age could be one explanation. To address this, Figure 1 also gathers comparable estimates from children closer to PIRLS age. Although some are from small or nonrepresentative samples, they demonstrate that higher agreement on other measures is not confined to PISA.[7]

[6] Moreover, students are are asked to exclude "school books" and given visual cues for each category, not reproduced here, while parents are not.

[7] Reported are the median estimates from West, Sweeting, and Speed (2001) and Vereecken and Vandegehuchte (2003) for occupation, Andersen et al. (2008) for family affluence, and Ensminger et al. (2000) for education. Family affluence is a summed index comprising the number of cars, computers and family vacations, and whether the respondent has their own bedroom. Estimates for family affluence refers to weighted κ and so are artificially somewhat higher. The full range of estimates are .57–.72 in West et al. (N=1267–1476), .58–.76 in Vereecken and Vandegehuchte (N=200), .43–.63 in Ensminger et al. (N=119), and .34–.63 in Andersen et al. (N=915).

Another concern is that the items asked are not identical. Therefore, I estimate a total number of books from the parent questionnaire by addition of midpoints (e.g., "101–200 books" and "51–100 children's books" will sum to ">200 books" as 150 + 75 = 225). This should improve agreement if questionnaire design were to explain the lack of it. Instead κ actually deteriorates somewhat, suggesting an explanation has to be sought elsewhere. Finally, Figure 1, right panel, displays rank order correlations. This is a more appropriate metric for the children's books item where the categories are not comparable, and could also be important if students use a different factor to convert books into shelves than intended (see Table 1). These figures are higher, but still fall short of comparable estimates in the literature.[8] The upshot is that low agreement on NBH cannot be accounted for by questionnaire differences or student age.

B. The Structure of Disagreement

The κ statistics around .20 in Figure 1 translate into a percentage agreement of about 40%, implying that 60% report a different category than their parent. In fact, there is no single country where a majority of reports agree. The direction of this disagreement is of some interest because of its implications for bias. As discussed above, if underreporting is more common among low achievers, the importance of books for achievement would be overstated in regression analyses that will pick up a positive term $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$. To assess this, I estimate misreporting as the difference between student reports and the estimated total from parent reports (and ignore for now that this might also be endogenous due to a positive term $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$).

Using pooled data, Figure 2 shows the probability that the student reports a higher or lower category than the parent ('over' and 'underreport') by the parent's category and the student's decile in the national achievement distribution. Student overreporting is a relatively rare phenomenon, except when parents report in one of the bottom two categories. In contrast, underreporting is much more common. For students of median achievement whose parents report in the middle ("26-100 books"), the probability of an underreport outweighs that of an overreport by a factor of three (.455 vs. .146).

Importantly, underreporting is clearly associated with reading achievement while overreporting is not. Focusing again on students whose parents report in the middle category ("26-100 books"), moving from the top to the bottom of the achievement distribution increases the probability of an underreport by a factor of 1.6 (.567 vs. .351). This difference is even starker in the category below ("11-25 books") with a factor of nearly two (.407 vs. .219). Taking into account the extent of disagreement – the number of categories by which reports differ – accentuates these patterns even further (results not shown).

[8] Engzell and Jonsson (2015, p. 325) report Spearman's ρ from 14 year olds in the range of .41–.59 for parental education and .62–.74 for occupation (.32–.66 and .48–70 if the parent was foreign born). Cohen and Orum (1972) report γ correlations from 9–13 year olds of .62.–72 for education and .75–.85 for occupation. Andersen et al. (2008) report γ of .53–.80 on their family affluence scale.

C. *Learning from Gender Differences*

The above findings are suggestive of differential error but may be sensitive to the assumption that parents report correctly. They also obscure that parent reports likewise may be subject to endogeneity. While it is plausible to think that parents' reporting error, if any, does not depend on student achievement (so that $\boldsymbol{\eta} \perp \boldsymbol{\varepsilon}$), if there is endogeneity in actual books this would affect both sources through the term $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$. As a way to test either type of endogeneity, I turn to an exogenous source of achievement: student's gender.[9] Because this strategy does not rely on linking sources I am also able to examine PISA data, where parents are not asked about NBH.

Parents do not usually choose the gender of their child but girls outperform boys in reading throughout the school age. The reasons for this gap are disputed, but it is clear that it opens up at a very early age, and likely has some biological underpinnings (Baker and Milligan 2013). Mullis et al. (2012) study these differences in PIRLS 2011 and report a sizeable female advantage that is statistically significant in all but five countries. Similar results for PISA 2012 are reported in OECD (2015). Following the above we would expect girls and, possibly, their parents to report higher NBH.

The results, reported as odds ratios in Figure 3, are striking. In both PISA and PIRLS, girls tend to report a higher number of books, sometimes by a wide margin. The gender difference is consistent between the two surveys, but varies across countries. Differences in parent reports are of a lesser magnitude and confined to children's books. This supports the interpretation that student reports are twice endogenous, due both to endogenous inputs $\boldsymbol{\xi}'\boldsymbol{\varepsilon}$ and differential misreporting $\boldsymbol{\eta}'\boldsymbol{\varepsilon}$. Finally, there is an opposite gender difference for student reports on parents' education. This is consistent with the notion that low achievers are prone to exaggerate this variable, which would bias regression estimates further downward compared to the classical case.

V. A Bias Decomposition

Validation studies of survey reported data have paid considerable attention to the possibility of errors being correlated with true values, demographic characteristics, or across time (Bound et al. 2001). Attempts to test for differentiality are more rare, but an exception is Black, Sanders, and Taylor (2003) who consider errors in Census reported education in a model of earnings determination. Following Bound et al. (1994), they show how to decompose bias from

[9] Little evidence exists for a systematic correlation between parent ability or status and offspring gender; see the extensive discussion by Kolk and Schnettler (2013) and references therein. The leading theory in favour of one predicts male (female) biased sex ratios for high (low) status parents (Almond and Edlund 2007), which would bias the gender difference in reported NBH documented here toward zero. A related worry is that girl children might influence assets through a detrimental effect on marital stability, as suggested by Mammen (2008) and others. There is little evidence to sustain this beyond the U.S. case (Diekmann and Schmidheiny 2004), and again, the hypothesized direction would bias the differential in NBH toward zero.

arbitrary error into attenuation and endogeneity. Their method builds on the strong assumption that validation data represent the truth, which, as Abowd and Stinson (2013) stress, is usually untenable. A further complication arises in our case due to the composite nature of endogeneity: ideally we would like to distinguish endogenous misreporting from endogeneity of inputs.

To see how these issues can be dealt with, it is helpful to first describe the approach of Black et al. (2003). For reasons given above, data collected from parents are generally assumed more reliable. The absence of gender differences also demonstrates that parent reports about non-children's books are the only information on NBH that is rid of endogeneity, so it is natural to take this variable as a benchmark for how well we can reasonably hope to measure the variable. Denote the student and parent reports $\boldsymbol{M}_s$ and $\boldsymbol{M}_p$, and assume $\boldsymbol{M}_p - \boldsymbol{X} = \boldsymbol{0}$. Under the strong assumption of no error in parental data, we can consistently estimate the error, coefficient vector, and residuals as:

$$\begin{aligned} \widehat{\boldsymbol{U}} &= \boldsymbol{M}_s - \boldsymbol{M}_p \\ \widehat{\boldsymbol{\beta}} &= \boldsymbol{\Sigma}_{M_p}^{-1} \boldsymbol{M}_p' \boldsymbol{y} \\ \widehat{\boldsymbol{\varepsilon}} &= \boldsymbol{y} - \boldsymbol{M}_p \widehat{\boldsymbol{\beta}} \end{aligned}$$

Substituting these into equation (3) produces the following least squares decomposition that Black et al. (2003) worked with:

$$\underbrace{\boldsymbol{b}_s - \widehat{\boldsymbol{\beta}}}_{\text{bias}} = \underbrace{\boldsymbol{\Sigma}_{M_s}^{-1} \boldsymbol{M}_s' \widehat{\boldsymbol{U}}(-\widehat{\boldsymbol{\beta}})}_{\text{attenuation}} + \underbrace{\boldsymbol{\Sigma}_{M_s}^{-1} \boldsymbol{M}_s' \widehat{\boldsymbol{\varepsilon}}}_{\text{endogeneity}} \tag{4}$$

where $\boldsymbol{b}_s$ is the naive slope estimate obtained from mismeasured variables, in our case the student data.

Particular to our application is that the endogenous component consists of part differential misreporting ($\boldsymbol{\eta}'\boldsymbol{\varepsilon}$), part reciprocal causation due to endogeneity of inputs ($\boldsymbol{\xi}'\boldsymbol{\varepsilon}$). In the above decomposition, both get absorbed into the last term. To approximate the relative contribution of each, we can use the fact that PIRLS asks parents a separate question about children's books. The key assumption will be that $\boldsymbol{M}_s$ does not contain any information about $\boldsymbol{\xi}$, the stochastic component of books, once conditioning on children's books. Write the residuals from a regression of student reports on this variable $\boldsymbol{M}_s^*$. Then:

$$\underbrace{\boldsymbol{b}_{\widehat{\varepsilon} M_s}}_{\text{endogeneity}} = \underbrace{\boldsymbol{b}_{\widehat{\varepsilon} M_s} - \boldsymbol{\Sigma}_{M_s^*}^{-1} \boldsymbol{M}_s^{*\prime} \widehat{\boldsymbol{\varepsilon}}}_{\text{reciprocal causation}} + \underbrace{\boldsymbol{\Sigma}_{M_s^*}^{-1} \boldsymbol{M}_s^{*\prime} \widehat{\boldsymbol{\varepsilon}}}_{\text{misreporting}} \tag{5}$$

with $\boldsymbol{b}_{\widehat{\varepsilon} M_s}$ used as a shorthand for the last term of equation (4). Thus we are assuming that once $\boldsymbol{M}_s$ has been purged of any variation explained by (parent reported) children's books, the remaining association with residuals $\widehat{\boldsymbol{\varepsilon}}$ can be interpreted as being due to differential misreporting, $\boldsymbol{\eta}$.

This might be a strong assumption, for two reasons. One problem is that measurement error in children's books could understate the contribution of endogenous inputs, but this is addressed in the sensitivity analysis to follow. The other problem is that children's books might be reflective of parents' characteristics and not just the child's achievement. This bias will work in the opposite direction, potentially overstating the contribution of inputs. If the latter seems to account for most endogeneity, therefore, this cannot be taken as conclusive evidence that differential misreporting is unimportant. On the other hand, if inputs fail to account for much of the endogenous bias throughout the specifications, the implication would seem to be that differential misreporting is of consequence.

A. Allowing for Error in Parent Reports

To assess robustness to errors in parent reported data, I employ a version of the simulation–extrapolation or *simex* algorithm for regression deconvolution introduced by Cook and Stefanski (1994) and adapted for categorical data by Küchenhoff, Mwalili, and Lesaffre (2006). This approach is potentially useful when both sources may be subject to non-classical disturbances, but one of these can credibly be described as a Monte Carlo process using a limited set of parameters. A general motivation of the method is provided by Carroll et al. (2006), who discuss its theoretical affinities with jackknife estimation and equivalence to the method of moments in simpler linear settings. Despite a wide range of recent applications (e.g., Delaigle, Hall, and Jamshidi 2015, Lockwood and McCaffrey 2015), it has to my knowledge not been applied to the problem of error in validation data.

As a Monte Carlo based estimator, *simex* makes minimal parametric assumptions. In particular, it does not impose any distribution on the unobserved regressor(s) $\boldsymbol{X}$ and allows us to maintain a fully arbitrary error structure in student reports, which is essential. The tradeoff is that it requires an explicit specification of the error in the validation data. Here we are guided by prior knowledge: analyses by student gender demonstrate that this error is nondifferential, and it is necessarily categorical and bounded because the variable itself is. Nevertheless, without direct evidence about parents' reliability, some guesswork is inevitable. I therefore simulate several scenarios, to be interpreted as a sensitivity analysis in the spirit of Rosenbaum and Rubin (1983), Horowitz and Manski (1995), Hyslop and Imbens (2001), Bollinger (2003), or Kreider (2010).[10]

Hypothetical error in parents is described as a 5×5 matrix $\mathbf{\Pi}$ where element π_{ij} states the probability of reporting in category i given true unobserved value j. Three scenarios are assessed, letting 10%, 20%, and 30% of parents misreport.[11] The idea is then to: (1) reestimate the model

[10] The approach taken here differs from some of these in that, rather than deriving analytical bounds under a 'worst case' scenario, it imposes a more realistic error structure and aims to establish probability limits by Monte Carlo simulation.

[11] To structure the off-diagonals, I take errors by $c + 1$ categories to be half as common as by c ($\geqslant 1$) categories within the bounds of the variable, that is, larger deviations are assumed less likely. Assuming

under repeated simulation of added noise, (2) fit a regression curve to parameter decay as a function of contamination, (3) extrapolate this curve to the ideal case of no error. For a given error structure, asymptotically unbiased estimates will be recovered as long as parameter decay is a continuous function of incremental noise and the extrapolation function is correctly specified.

Concretely, pseudo data are generated as random draws from the observed data subject to conditional probability $\mathbf{\Pi}^{\lambda}$, for successive contamination levels λ fixed on an equidistant grid $\{0, .25, .5, \ldots, 2\}$ and powers of $\mathbf{\Pi}$ obtained via the eigendecomposition $\mathbf{\Pi}^{\lambda} = \boldsymbol{V}\boldsymbol{D}^{\lambda}\boldsymbol{V}^{-1}$. At each level, $B = 50$ Monte Carlo draws are made, the decomposition reestimated, and an average of the B estimates of each parameter θ is computed as $\widehat{\theta}_{\lambda}$. A trend of bias is then established by fitting a parametric function $\widehat{\theta}_{\lambda} = g(\lambda)$, here a quadratic polynomial: $\widehat{\theta} = \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2$. The last step extrapolates this function to $\widehat{\theta}_{-1}$, where parameter decay has been 'reversed', figuratively speaking. Operational choices described in this paragraph are conventional and results are sensitive mainly to the specification of the original input matrix, $\mathbf{\Pi}$ (cf. Carroll et al. 2006, Küchenhoff et al. 2006).

B. Decomposition Results

To illustrate the bias in a simple setting, I estimate a bivariate regression separately for each of the 40 countries in the sample.[12] Similar cross-country comparisons are presented by Martins and Veiga (2010), Raitano and Vona (2016), or Schütz et al. (2008). Like the studies by Schütz et al. (2008) and Ammermueller and Pischke (2009), I enter NBH as a continuous variable with range 1–5, letting the slope estimate reflect the advantage in reading associated with one step up the 'ladder' of categories. Compared to a categorical (dummy variable) specification, this simplifies the exposition without altering any substantive conclusions. An exception is made for the estimation of residual variation in student reports $\boldsymbol{M}_s^*$, where validation data on children's books are entered categorically to make maximal use of the information contained therein.

I first focus on the limiting case of no error in the validation data, where Table 2 shows point estimates from the decomposition. For countries with low numbers of books, toward the bottom of the table, parent reports yield larger estimates. In countries where aggregate numbers are higher, and therefore, the scope for endogenous underreporting larger, there is less of a consistent pattern: student reports yield estimates that are variously smaller, larger, or of comparable size.

random misclassification naturally causes more rapid parameter decay. I also allow parents' errors to the two questions to be correlated ($r \approx .3$) following reasoning in Lubotsky and Wittenberg (2006). Taking errors to be orthogonal, the role of endogenous inputs remains minor across all specifications but other results remain unchanged. Conversely, allowing maximally correlated errors increases the contribution of endogenous inputs. Lastly, I assume that parents' error is generated independently of the student's. This assumption is not so much empirical as it is conceptual: to the extent that correlated errors exist, these are likely to reflect durable and transmitted attitudes that rather belong in a definition of the target construct.

[12] These results do not change appreciably with the inclusion of controls for gender, age, or foreign language use at home.

Of the bias components, endogeneity tends be more variable across countries than attenuation: the standard deviation of these two statistics across countries is .023 and .050, respectively. The rightmost two columns reveal that differential misreporting accounts for most of the endogenous bias in this specification, about three fourths in the median country, and reciprocal causation only a minor part.

The next question is what happens once we relax the assumption of no errors in the validation data; results are found in Figure 4. The solid line plots the distribution of country estimates assuming that parents are correct and conveys the same information as Table 2. Dashed lines show how each set of estimates changes once we allow for the possibility of increased error among parents: 10%, 20%, and 30% misclassified. The most obvious consequence is that $\widehat{\boldsymbol{\beta}}$ increases, which shifts the total bias downwards (top, left). At the same time, the variance of $\widehat{\boldsymbol{U}}$ decreases and these two have offsetting effects on estimated attenuation (top, right). The shift in total bias is instead driven by a corresponding decline in the endogenous component (bottom, left). This, in turn, is somewhat offset by an increased role for reciprocal causation (bottom, right). The latter fails to account for much of the endogenous bias until we allow an error rate of 20% where it contributes on average about half, while at 30% the balance tips the other way.

Which, if any, of these scenarios is most plausible? Unfortunately there are no data that allow us to assess the reliability of parents, but the answer depends in large part on what we take the proxy to reflect. If we want to learn about the actual number of books, parents are likely to report with considerable error and 30% may be closer to the truth. However, a more common interpretation is in terms of underlying characteristics such as "whether the parents value literary skills" (Ammermueller and Pischke 2009, p. 322). In this case, parent reports would seem true as a matter of definition, save for chance fluctuation in what would be answered from one occasion to the next. If so, an error estimate of 10% might be more appropriate.

To make intuitive sense of results in a comparative setting, Figure 5 plots the estimate from student reports along with the separate bias components under the (perhaps conservative) scenario of 10% errors in validation data. Dashed lines mark the 10 to 90 percentile range of each bias component. This graph reaffirms the impression that attenuation does not vary markedly across countries. If this was the only source of bias, the impact of family background would be underestimated but about equally much so in all countries. Once we turn to the endogenous component, the picture changes. The variability is larger and clearly contributes to the substantive estimate, so that the cross-country pattern is attributable in part to the extent of endogeneity. Admittedly, the relative balance of these components shifts once we allow for larger errors in validation data, but endogeneity remains important in all of the specifications above.

C. *Relation to Gender Differences*

While the decomposition results inevitably depend on assumptions about the error in validation data, gender differences provide more indisputable evidence of endogeneity. A question that we have not asked so far is whether the pattern of endogeneity across countries is consistent depending on whether gender differences or validation reports are used as a guide. One could imagine various ways to go about this, but one way is to run a countrywise regression of student reported values $\boldsymbol{M}_s$ on the residuals $\hat{\varepsilon}$ and multiply this by the estimated gender difference in achievement. This retrieves the expected gender difference in reported categories based on the decomposition, assuming linearity and that gender fulfils the criterion of a valid instrument for achievement (i.e., impacts reported books only through its predictive power for achievement as measured in PIRLS).

Gender is, however, unlikely to satisfy this exclusion restriction. Among other things, gendered parenting practices and the limits of test scores as a measure of achievement entail that the differential in reported books may outsize the model prediction. Indeed, this turns out to be the case by a factor of 2.8 in the median country. But even if absolute comparison of these coefficients is not informative, the relative pattern across countries might be. The cross-country correlation of the two sets of estimates is r=.535 (40 countries). Excluding countries where the majority of students attend single sex schools (Iran, Qatar, Saudi Arabia, United Arab Emirates) which could bias recruitment into the sample, the correlation rises to r=.688 (36 countries). Corresponding rank order correlations are ρ=.592 and ρ=.705. Given the limits of gender differences as a means of quantifying endogeneity, these results seem to confirm those from the decomposition.

VI. Conclusion

As a proxy for student background, self-reported books in the home are subject to sizeable and systematic errors of observation. Not only do students from bookish homes perform better, but better students also accumulate more books and are better informed about their home libraries. The resulting endogenous bias leads to associations of an impressive size. But these do not have a meaningful interpretation in terms of the variable's reliability or substantive importance compared to other proxies, and ultimately researchers may be better served by proxies that evince more modest associations.

This example serves as a useful reminder of the caution by Bound et al. (2001, p. 3709) that "non-classical measurement error should be taken much more seriously by those who analyze survey data" (cf. Kreider 2010). In the decade-and-half intervening since their writing, much has been done to generalize the classical model beyond the standard linear case (e.g., Chen et al. 2011, Battistin and Sianesi 2011), while the assumption of nondifferentiality continues to be

invoked as a matter of convention. This is understandable given the inferential obstacles raised by violations of it, but it places a responsibility on applied researchers to proceed with caution.

In a comparative setting, endogenous bias in NBH appears to be more variable than attenuation, distorting cross-country patterns. One source of variation is the aggregate distribution: in countries where many books are the norm, the scope for endogenous underreporting is larger. Endogeneity also entails that any increase in the *variance* of achievement will inevitably lead to the impression of an increased family background association. This is perhaps most consequential in designs that attempt to control for unobserved heterogeneity at the country level (e.g., fixed effects or differences-in-differences), which become vulnerable to spurious results because true variation in the underlying association is smaller. Other questions addressed in this literature include whether socioeconomic gaps vary by student age, student gender, or achievement domain such as reading or mathematics. The above analysis suggests that NBH is ill suited as a proxy in each of these cases, as it is plausible that endogeneity differs along several or all of these dimensions. These findings add to an expanding list of difficulties in making policy relevant inferences from student achievement data (Bond and Lang 2013, Kreiner and Christensen 2014, Contini and Grand 2015).

An additional implication concerns models that attempt to identify peer influence on achievement. Researchers since Manski (1993) have been aware of the difficulties confronting the estimation of endogenous (i.e., achievement-on-achievement) effects and resorted to a reduced-form specification with peer composition defined by predetermined characteristics. As an endogenous variable, self-report NBH does not belong in this category and if used risks reintroducing Manski's reflection problem through the back door.

These issues are likely to be exacerbated when attempts at bias correction are made that rely on classical assumptions. For example, Ammermueller and Pischke (2009) instrument parent reported NBH with student reports to compensate for attenuation, as they note is standard when separate reports by two individuals are available. Knowledge of endogeneity here suggests that 'corrected' estimates are probably greatly exaggerated, and 'uncorrected' estimates closer to the truth. Using parent rather than student reports as the instrument is no remedy in this case: the error in student reports is both endogenous and mean reverting, entailing upward bias in either case (Kane et al. 1999).

The problems uncovered here are perhaps especially troubling in school surveys, where the same competencies that determine response quality appear as a test score at the left-hand side of the equation. But similar problems arguably threaten to arise in various other areas of survey research. Retrospective reports about childhood NBH have been collected in surveys of adult or elderly respondents in recent years (e.g., Brunello et al. 2015). An interesting question beyond the scope of this paper is whether these reports are subject to similar biases, but it is not unreasonable to think that they might be.

Lastly, the presence of large, differential error documented here is of some relevance for a recent argument about how to extract information from several proxies, as raised by Lubotsky and Wittenberg (2006). They show how, under nondifferentiality, partial coefficients from a multivariate regression can be aggregated so as to minimize attenuation. The problem is that this assumption will rarely if ever be tested, but violations are arguably to be expected when a large number of noisy measures are used. The more standard practice of extracting a common factor among proxies in part alleviates this problem, but potentially at the cost of introducing errors that are correlated across them. Ultimately the issue comes down to choosing among evils, and which strategy is the most credible must be decided from the case at hand. Further investigation of this issue across the relevant empirical contexts would be valuable.

References

Abowd, John M., and Martha H. Stinson (2013). "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," Review of Economics and Statistics, 95(5), 1451-1467.

Algan, Yann, Pierre Cahuc, and Andrei Shleifer (2011). "Teaching Practices and Social Capital (No. 17527)," Cambridge, MA: NBER.

Algan, Yann, Pierre Cahuc, and Andrei Shleifer (2013). "Teaching Practices and Social Capital," American Economic Journal: Applied Economics, 5(3), 189-210.

Almond, Douglas, and Lena Edlund (2007). "Trivers–Willard at Birth and One Year: Evidence From US Natality Data 1983-2001," Proceedings of the Royal Society B: Biological Sciences, 274(1624), 2491-2496.

Ammermueller, Andreas (2007). "PISA: What Makes the Difference? Explaining the Gap in PISA Test Scores Between Finland and Germany," Empirical Economics, 32(2), 263-287.

Ammermueller, Andreas (2013). "Institutional Features of Schooling Systems and Educational Inequality: Cross-Country Evidence From PIRLS and PISA," German Economic Review, 14(2), 190-213.

Ammermueller, Andreas, and Jörn-Steffen Pischke (2009). "Peer Effects in European Primary Schools: Evidence From the Progress in International Reading Literacy Study," Journal of Labor Economics, 27(3), 315-348.

Andersen, Anette, Rikke Krølner, Candace Currie, Lorenza Dallago, Pernille Due, Matthias Richter, Agota Örkényi, and Bjørn Evald Holstein (2008). "High Agreement on Family Affluence Between Children's and Parents' Reports," Journal of Epidemiology and Community Health, 62(12), 1092-1094.

Baker, Michael, and Kevin Milligan (2013). "Boy-Girl Differences in Parental Time Inputs: Evidence From Three Countries (No. 18893)," Cambridge, MA: NBER.

Battistin, Erich, and Barbara Sianesi (2011). "Misclassified Treatment Status and Treatment Effects: An Application to Returns to Education in the United Kingdom," Review of Economics and Statistics, 93(2), 495-509.

Betts, Julian R. (2011). "The Economics of Tracking in Education," in Hanushek, E. A., Machin, S., Woessmann, L. (Eds.): Handbook of the Economics of Education, 3, 341-381. Amsterdam: North-Holland.

Björklund, Anders, and Kjell G. Salvanes (2011). "Education and Family Background: Mechanisms and Policies," in Hanushek, E. A., Machin, S., Woessmann, L. (Eds.): Handbook of the Economics of Education, 3, 201-247. Amsterdam: North-Holland.

Black, Dan, and Jeffrey A. Smith (2006). "Estimating the Returns to College Quality with Multiple Proxies for Quality," Journal of Labor Economics, 24(3), 701-728.

Black, Dan, Seth Sanders, and Lowell Taylor (2003). "Measurement of Higher Education in the Census and Current Population Survey," Journal of the American Statistical Association, 98(463), 545-554.

Bollinger, Christopher R. (2003). "Measurement Error in Human Capital and the Black-White Wage Gap," Review of Economics and Statistics, 85(3), 578-585.

Bond, Timothy N., and Kevin Lang. (2013). "The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results." Review of Economics and Statistics 95(5), 1468-1479.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001). "Measurement Error in Survey Data," in Heckman, J. J., Leamer, E. E. (Eds.): Handbook of Econometrics, 5, 3705-3843. Amsterdam: North-Holland.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers (1994). "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," Journal of Labor Economics, 345-368.

Brese, Falk, and Plamen Mirazchiyski (2013). Measuring Students' Family Background in Large-Scale International Education Studies. Hamburg: IEA-ETS Research Institute.

Brunello, Giorgio, and Daniele Checchi (2007). "Does School Tracking Affect Equality of Opportunity? New International Evidence," Economic Policy, 22(52), 782-861.

Brunello, Giorgio, Guglielmo Weber, and Christoph T. Weiss (2015). "Books Are Forever: Early Life Conditions, Education and Lifetime Earnings in Europe," Economic Journal, early access. doi: 10.1111/Ecoj.12307

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu (2006). Measurement Error in Nonlinear Models: A Modern Perspective. CRC Press.

Chen, Xiaohong, Han Hong, and Denis Nekipelov (2011). "Nonlinear Models of Measurement Errors," Journal of Economic Literature, 49(4), 901-937.

Cohen, Roberta S., and Anthony M. Orum (1972). "Parent-Child Consensus on Socioeconomic Data Obtained From Sample Surveys," Public Opinion Quarterly, 36(1), 95-98.

Contini, Dalit, and Elisa Grand. (2015). "On Estimating Achievement Dynamic Models from Repeated Cross Sections," Sociological Methods & Research, early access. doi: 10.1177/0049124115613773

Cook, John R., and Leonard A. Stefanski (1994). "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," Journal of the American Statistical Association, 89(428), 1314-1328.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010). "Estimating the Technology of Cognitive and Noncognitive Skill Formation," Econometrica, 78(3), 883-931.

De Witte, Kristof, and Mika Kortelainen (2013). "What Explains the Performance of Students in a Heterogeneous Environment?," Applied Economics, 45(17), 2401-2412.

Delaigle, Aurore, Peter Hall, and Farshid Jamshidi (2015). "Confidence Bands in Non-Parametric Errors-in-Variables Regression," Journal of the Royal Statistical Society, Series B, 77(1), 149-169.

Diekmann, Andreas, and Kurt Schmidheiny (2004). "Do Parents of Girls Have a Higher Risk of Divorce? An Eighteen-Country Study," Journal of Marriage and Family, 66(3), 651-660.

Engzell, Per, and Jan O. Jonsson (2015). "Estimating Social and Ethnic Inequality in School Surveys: Biases From Child Misreporting and Parent Nonresponse," European Sociological Review 31(3), 312-325.

Ensminger, Margaret E., Christopher B. Forrest, Anne W. Riley, Myungsa Kang, Bert F. Green, Barbara Starfield, and Sheryl A. Ryan (2000). "The Validity of Measures of Socioeconomic Status of Adolescents," Journal of Adolescent Research, 15(3), 392-419.

Epple, Dennis, and Richard Romano (2011). "Peer Effects in Education: A Survey of the Theory and Evidence," in Benhabib, J., Bisin, A., and Jackson, M. O. (Eds.): Handbook of Social Economics, 1, 1053-1163. Amsterdam: North-Holland.

Ferreira, Francisco H. G., and Jérémie Gignoux (2014). "The Measurement of Educational Inequality: Achievement and Opportunity," World Bank Economic Review, 28(2), 210-246.

Freeman, Richard B., and Martina Viarengo (2014). "School and Family Effects on Educational Outcomes Across Countries," Economic Policy, 29 (79), 395-446.

Freeman, Richard B., Stephen Machin, and Martina Viarengo (2010). "Variation in Educational Outcomes and Policies Across Countries and of Schools Within Countries (No. 16293)," Cambridge, MA: NBER.

Fryer Jr, Roland G., and Steven D. Levitt (2004). "Understanding the Black-White Test Score Gap in the First Two Years of School," Review of Economics and Statistics, 86(2), 447-464.

Fuchs, Thomas, and Ludger Woessmann (2007). "What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data," Empirical Economics, 32(2), 433-464.

Griliches, Zvi (1986). "Economic Data Issues," in Griliches, Z., Intriligator, M. D. (Eds.): Handbook of Econometrics, 3, 1465-1514. Amsterdam: North-Holland.

Hanushek, Eric A., and Ludger Woessmann (2011). "The Economics of International Differences in Educational Achievement,"in Hanushek, E. A., Machin, S., Woessmann, L. (Eds.): Handbook of the Economics of Education, 3, 89-200. Amsterdam: North-Holland.

Heckman, James J., and Stefano Mosso (2014). "The Economics of Human Development and Social Mobility," Annual Review of Economics, 6, 689-733.

Holley, Charles E. (1916). The Relationship Between Persistence in School and Home Conditions. University of Chicago Press.

Horowitz, Joel L., and Charles F. Manski (1995). "Identification and Robustness with Contaminated and Corrupted Data," Econometrica, 63(1), 281–302.

Hyslop, Dean R., and Guido W. Imbens (2001). "Bias From Classical and Other Forms of Measurement Error," Journal of Business and Economic Statistics, 19(4), 475-481.

Jerrim, John, and John Micklewright (2014). "Socio-Economic Gradients in Children's Cognitive Skills: Are Cross-Country Comparisons Robust to Who Reports Family Background?," European Sociological Review, 30(6), 766-781.

Jürges, Hendrik, and Kerstin Schneider (2007). "Fair Ranking of Teachers," Empirical Economics, 32(2), 411-431.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel (2005). "The Effect of Central Exit Examinations on Student Achievement," Journal of the European Economic Association, 3(5), 1134-1155.

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger (1999). "Estimating Returns to Schooling When Schooling Is Misreported (No. 7235)," Cambridge, MA: NBER.

Kerckhoff, Alan C., William M. Mason, and Sharon S. Poss (1973). "On the Accuracy of Children's Reports of Family Social Status," Sociology of Education, 46, 219-247.

Kolk, Martin, and Sebastian Schnettler (2013). "Parental Status and Gender Preferences for Children: Is Differential Fertility Stopping Consistent with the Trivers–Willard Hypothesis?," Journal of Biosocial Science, 45(05), 683-704.

Kreider, Brent (2010). "Regression Coefficient Identification Decay in the Presence of Infrequent Classification Errors," Review of Economics and Statistics, 93(4), 1017-1023.

Kreiner, Svend, and Karl Bang Christensen. (2014). "Analyses of Model Fit and Robustness: A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy," Psychometrika 79(2), 210-231.

Kreuter, Frauke, Stephanie Eckman, Kai Maaz, and Rainer Watermann (2010). "Children's Reports of Parents' Education Level: Does It Matter Whom You Ask and What You Ask About?," Survey Research Methods, 4(3), 127-138.

Küchenhoff, Helmut, Samuel M. Mwalili, and Emmanuel Lesaffre (2006). "A General Method for Dealing with Misclassification in Regression: the Misclassification SIMEX," Biometrics, 62(1), 85-96.

Landis, J. Richard, and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data," Biometrics, 33(1), 159-174.

Leamer, Edward E. (1983). "Model Choice and Specification Analysis," in Intriligator, M. D., Griliches, Z. (Eds.): Handbook of Econometrics, 1, 285-330. Amsterdam: North-Holland.

Lockwood, J. R., and McCaffrey, Daniel F. (2015). "Matching and Weighting with Functions of Error-Prone Covariates for Causal Inference," Journal of the American Statistical Association, early access. doi: 10.1080/01621459.2015.1122601

Looker, E. Dianne (1989). "Accuracy of Proxy Reports of Parental Status Characteristics," Sociology of Education, 62, 257-276.

Lubotsky, Darren, and Martin Wittenberg (2006). "Interpretation of Regressions with Multiple Proxies," Review of Economics and Statistics, 88(3), 549-562.

Mammen, Kristin (2008). "The Effect of Children's Gender on Living Arrangements and Child Support," American Economic Review, 98(2), 408-412.

Manski, Charles F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem," The Review Of Economic Studies 60(3), 531-542.

Manski, Charles F., and Elie Tamer (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome," Econometrica, 70(2), 519-546.

Martins, Lurdes, and Paula Veiga (2010). "Do Inequalities in Parents' Education Play an Important Role in PISA Students' Mathematics Achievement Test Score Disparities?," Economics of Education Review, 29(6), 1016-1033.

Mason, William M., Robert M. Hauser, Alan C. Kerckhoff, Sharon S. Poss, and Kenneth Manton (1976). "Models of Response Error in Student Reports of Parental Socioeconomic Characteristics," in W. H. Sewell, R. M. Hauser, D. L. Featherman (Eds.), Schooling and Achievement in American Society, 443-494. New York: Academic Press.

Micklewright, John, and Schnepf, Sylke V. (2010). "How Reliable Are Income Data Collected with a Single Question?," Journal of the Royal Statistical Society, Series A, 173(2), 409-429.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Kathleen T. Drucker (2012). PIRLS 2011 International Results in Reading. Amsterdam: IEA.

Mullis, Ina V. S., Michael O. Martin, Ann M. Kennedy, Kathleen L. Trong, and Marian Sainsbury (2009). PIRLS 2011 Assessment Framework. Amsterdam: IEA.

New York Times, The (2011). "A Book in Every Home, and Then Some," The Opinion Pages, David Bornstein. May 16, 2011.

New York Times, The (2015a). "America's Students Are Lagging. Maybe It's Not the Schools," Economic Scene, Eduardo Porter. Nov 4, 2015, B1.

New York Times, The (2015b). "Our (Bare) Shelves, Our Selves," Future Tense, Teddy Wayne. Dec, 6, 2015, ST2.

OECD (2015). The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence. Paris: Organisation for Economic Co-Operation and Development.

Ohinata, Asako, and Jan C. van Ours (2013). "How Immigrant Children Affect the Academic Achievement of Native Dutch Children," Economic Journal, 123(570), F308-F331.

Peterson Paul E., and Ludger Woessmann (2007). "Introduction: Schools and the Equal Opportunity Problem," in Woessmann, L., Peterson, P. E. (Eds.), Schools and the Equal Opportunity Problem, 3-27," Cambridge, MA: MIT Press.

Raitano, Michele, and Francesco Vona (2013). "Peer Heterogeneity, School Tracking and Students' Performances: Evidence From PISA 2006," Applied Economics, 45(32), 4516-4532.

Raitano, Michele, and Francesco Vona (2016). "Assessing Students' Equality of Opportunity in OECD Countries: The Role of National-And School-Level Policies," Applied Economics, 1-16.

Rosenbaum, Paul R., and Donald B. Rubin (1983). "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," Journal of the Royal Statistical Society, Series B, 45(2), 212–218.

Rutkowski, Leslie, and David Rutkowski (2010). "Getting It Better: The Importance of Improving Background Questionnaires in International Large-Scale Assessments," Journal of Curriculum Studies, 42(3), 411-430.

Schneeweis, Nicole, and Rudolf Winter-Ebmer (2007). "Peer Effects in Austrian Schools," Empirical Economics, 32(2), 387-409.

Schütz, Gabriela, Heinrich W. Ursprung, and Ludger Woessmann (2008). "Education Policy and Equality of Opportunity," Kyklos, 61(2), 279-308.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge (2015). "What Are We Weighting For?," Journal of Human Resources, 50(2), 301-316.

Stapleton, Laura M., and Yoonjeong Kang (2016). "Design Effects of Multilevel Estimates From National Probability Samples," Sociological Methods and Research, early access. doi: 10.1177/0049124116630563

Thorndike, Robert L. (1973). "The Relation of School Achievement to Differences in the Backgrounds of Children," in Purves, A. C., Levine, D. U. (Eds.): Educational Policy and International Assessment: Implications of the IEA Surveys of Achievement, 93-103. Berkeley, CA: Mccutchan.

Traynor, Anne, and Tenko Raykov (2013). "Household Possessions Indices As Wealth Measures: A Validity Evaluation," Comparative Education Review, 57, 662-688.

Vereecken, C., and Vandegehuchte, A. (2003). "Measurement of Parental Occupation: Agreement Between Parents and Their Children," Archives of Public Health, 61, 141-149.

Waldinger, Fabian (2006). "Does Tracking Affect the Importance of Family Background on Students' Test Scores?," Mimeo, London School of Economics.

West, Patrick, Helen Sweeting, and Ewen Speed (2001). "We Really Do Know What You Do: A Comparison of Reports From 11 Year Olds and Their Parents in Respect of Parental Economic Activity and Occupation," Sociology, 35(2), 539-559.

Woessmann, Ludger (2003). "Schooling Resources, Educational Institutions and Student Performance: The International Evidence," Oxford Bulletin of Economics and Statistics, 65(2), 117-170.

Table 1: Books at home in the PIRLS 2011 student questionnaire, administered in school, and home questionnaire, distributed to student's parents or guardians. Adapted from original questionnaires available at: `http://timssandpirls.bc.edu/pirls2011/`.

Student questionnaire:

About how many books are there in your home? (Do not count magazines, newspapers, or your school books.) Fill one circle only.

None or very few (0–10 books) – ◯
Enough to fill one shelf (11–25 books) – ◯
Enough to fill one bookcase (26–100 books) – ◯
Enough to fill two bookcases (101–200 books) – ◯
Enough to fill three or more bookcases (more than 200) – ◯

Parent questionnaire:

About how many books are there in your home? (Do not count magazines, newspapers or children's books.) Check one circle only.

0–10 – ◯
11–25 – ◯
26–100 – ◯
101–200 – ◯
More than 200 – ◯

About how many children's books are there in your home? (Do not count children's magazines or school books.) Check one circle only.

0–10 – ◯
11–25 – ◯
26–50 – ◯
51–100 – ◯
More than 100 – ◯

Table 2: Estimates from bivariate linear regression of PIRLS 2011 reading scores on student and parent reports of number of books at home (range 1–5), and decomposition of the difference between the two assuming parent reports to be correct. 'Bias', 'Atten.', 'Endog.' refer to terms of equation (4), 'R. caus.' and 'Misrep.' to terms of equation (5). 95% confidence intervals adjusted for clustering at the school class level. Countries are ordered by average parent reported books, 'Median' refers to the median category reported by parents.

Country	N	Median	Student est.	Parent est.	Bias	Atten.	Endog.	R. caus.	Misrep.
Norway (nor)	2801	101–200	.246 (.018)	.265 (.019)	−.019	−.137	.118	.022	.096
Sweden (swe)	3837	101–200	.312 (.014)	.295 (.014)	.016	−.111	.128	.025	.103
Hungary (hun)	4832	26–100	.338 (.017)	.354 (.017)	−.015	−.117	.101	.031	.071
Denmark (dnk)	4299	26–100	.284 (.015)	.262 (.014)	.022	−.089	.111	.016	.094
Germany (deu)	2960	26–100	.315 (.017)	.302 (.016)	.013	−.137	.150	.028	.122
Georgia (geo)	4416	26–100	.186 (.014)	.243 (.019)	−.056	−.118	.061	.011	.050
Finland (fin)	4368	26–100	.271 (.016)	.241 (.014)	.030	−.108	.138	.032	.105
Austria (aut)	4356	26–100	.326 (.015)	.342 (.013)	−.016	−.135	.119	.037	.081
Czech Rep (cze)	4335	26–100	.339 (.016)	.276 (.015)	.063	−.120	.183	.019	.163
Canada (can)	18471	26–100	.246 (.008)	.171 (.007)	.075	−.094	.170	.010	.160
Ireland (irl)	4149	26–100	.339 (.014)	.280 (.013)	.059	−.124	.183	.036	.147
Malta (mlt)	3154	26–100	.219 (.020)	.215 (.016)	.004	−.137	.141	.040	.101
Spain (esp)	7827	26–100	.224 (.013)	.251 (.012)	−.026	−.118	.091	.025	.067
Russia (rus)	4399	26–100	.251 (.020)	.226 (.019)	.024	−.109	.134	.027	.107
Belgium Fr (bfr)	3300	26–100	.300 (.020)	.285 (.017)	.016	−.122	.138	.033	.105
France (fra)	4019	26–100	.311 (.016)	.273 (.015)	.038	−.116	.154	.034	.120
Slovakia (svk)	5414	26–100	.330 (.018)	.327 (.018)	.003	−.105	.108	.052	.056
Israel (isr)	3213	26–100	.198 (.020)	.291 (.019)	−.093	−.141	.048	.067	−.020
Bulgaria (bgr)	5041	26–100	.326 (.020)	.321 (.020)	.005	−.096	.101	.021	.081
Poland (pol)	4843	26–100	.295 (.015)	.282 (.013)	.013	−.137	.150	.027	.122
Italy (ita)	3806	26–100	.204 (.015)	.231 (.017)	−.027	−.107	.080	.028	.052
Slovenia (svn)	4274	26–100	.293 (.016)	.271 (.013)	.022	−.129	.151	.048	.103
Portugal (prt)	3845	26–100	.286 (.017)	.231 (.015)	.055	−.081	.136	.033	.102
Lithuania (ltu)	4367	26–100	.294 (.019)	.257 (.015)	.038	−.087	.125	.040	.084
Trinidad (tto)	3422	26–100	.176 (.020)	.222 (.019)	−.046	−.140	.094	.024	.070
Taiwan (twn)	4192	26–100	.233 (.013)	.206 (.013)	.027	−.089	.116	.010	.106
Singapore (sgp)	6077	26–100	.305 (.015)	.208 (.013)	.097	−.108	.205	.064	.142
Croatia (hrv)	4457	26–100	.230 (.016)	.250 (.014)	−.020	−.095	.075	.028	.047
Romania (rom)	4401	26–100	.347 (.021)	.340 (.018)	.006	−.097	.103	−.001	.104
Hong Kong (hkg)	3487	26–100	.123 (.020)	.112 (.017)	.011	−.055	.066	.025	.041
Qatar (qat)	3413	26–100	.024 (.016)	.180 (.020)	−.155	−.130	−.025	.030	−.055
UA Emirates (are)	12709	11–25	.134 (.013)	.237 (.013)	−.103	−.135	.032	.076	−.044
Saudi Arabia (sau)	4216	11–25	.082 (.022)	.150 (.021)	−.068	−.079	.011	.012	−.000
Oman (omn)	8752	11–25	.098 (.012)	.174 (.011)	−.076	−.114	.038	.024	.014
Azerbaijan (aze)	4272	11–25	.081 (.020)	.091 (.020)	−.010	−.062	.052	.005	.047
South Africa (zaf)	2605	11–25	.213 (.036)	.313 (.031)	−.100	−.145	.045	.027	.018
Iran (irn)	5515	11–25	.248 (.019)	.255 (.018)	−.007	−.140	.133	.047	.086
Colombia (col)	3669	11–25	.174 (.032)	.270 (.030)	−.097	−.147	.050	.017	.033
Morocco (mar)	5474	0–10	.119 (.024)	.158 (.023)	−.039	−.114	.076	−.001	.076
Indonesia (idn)	4400	0–10	.141 (.042)	.217 (.032)	−.075	−.147	.072	.006	.066

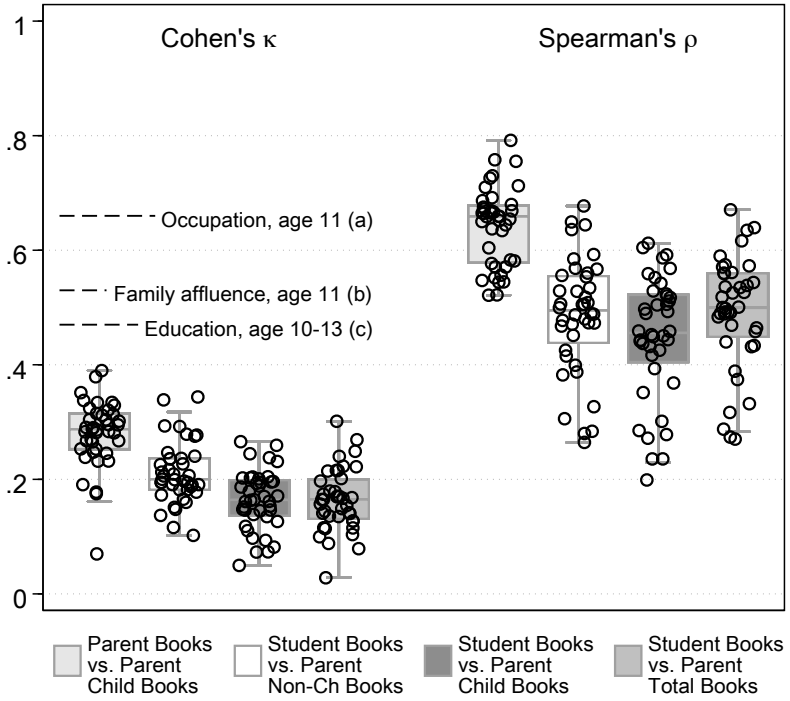

Figure 1: Agreement between students and parents on books in the home in PIRLS 2011. Cohen's κ (left) and Spearman's ρ (right). Each circle represents a country. The items are described in Table 1 and the running text. Median κ estimates from earlier studies are displayed for comparison (dashed lines), sources: (a) West et al. (2001), Vereecken and Vandegehuchte (2003), (b) Andersen et al. (2008), (c) Ensminger et al. (2000). N (PIRLS)=2,808–8,487 (per country), 197,387 (total).

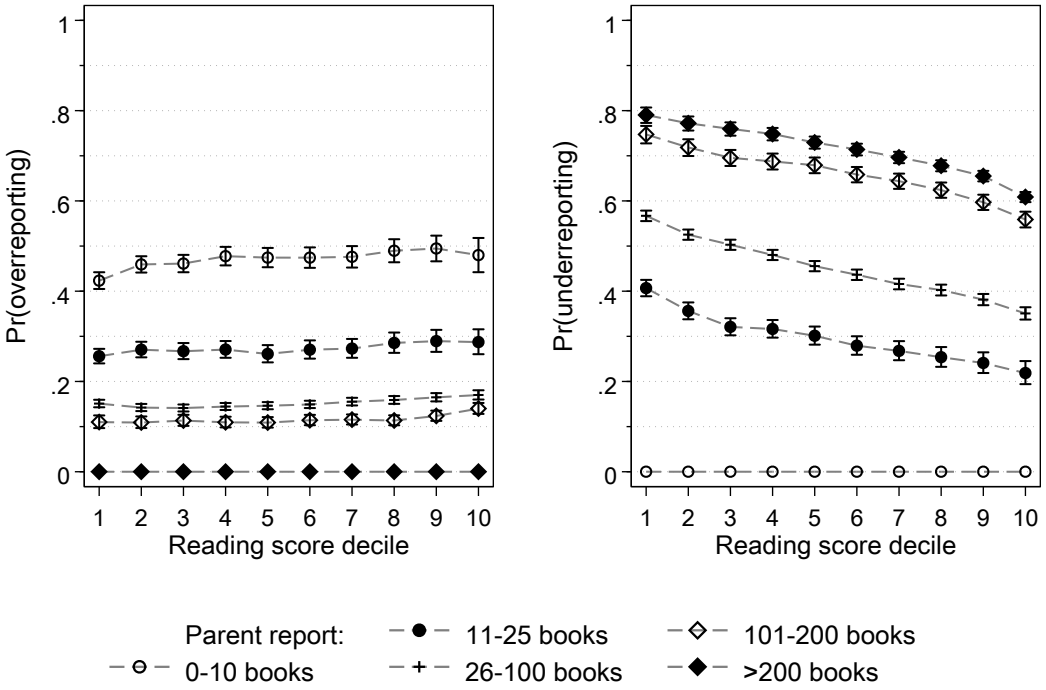

Figure 2: Estimated probability from fully interacted logistic regression of students reporting a higher or lower category than parent ('over' and 'underreporting'), by student's achievement decile and parent's reported value. Pooled data from PIRLS 2011, achievement scores standardized at the country level. 95% confidence intervals allowing for clustering on school classes. Underreporting is the most common form of disagreement, and closer associated with (low) achievement than overreporting. N=197,387.

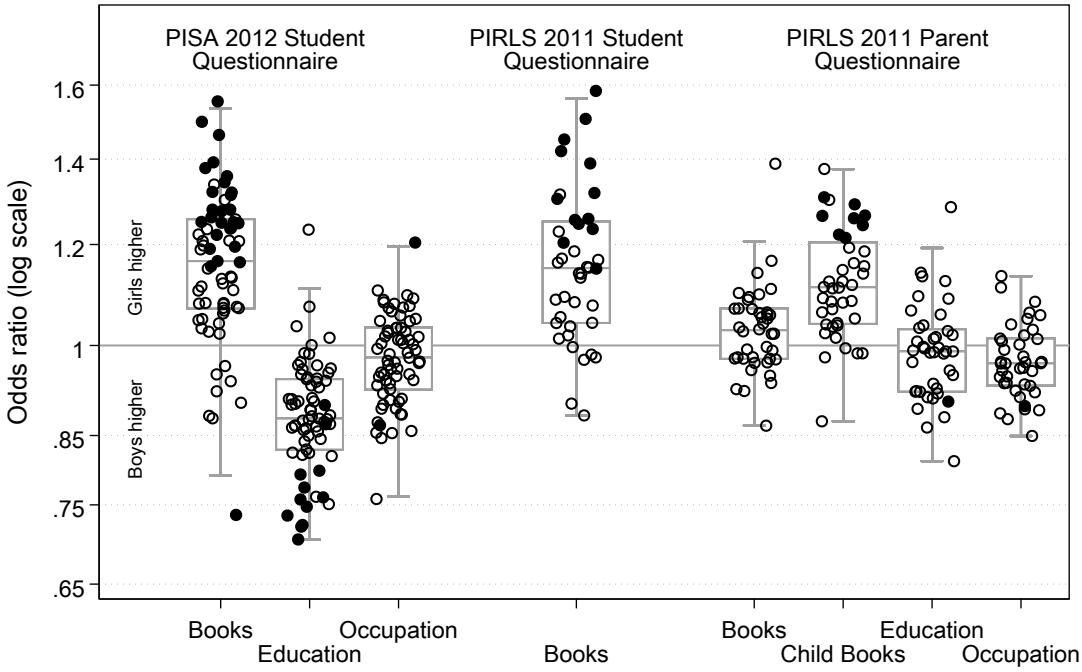

Figure 3: Odds ratios from ordered logistic regression of reported student background variables on student gender. Each circle represents a country. Filled markers indicate significance at the 5% confidence level, Bonferroni corrected by the number of study countries and allowing for clustering on the school (PISA) or school class (PIRLS) level. Higher values reported by girls (boys) are indicative of a positive (negative) endogenous bias. N (PISA)=1,334–28,074 (per country), 394,130 (total); N (PIRLS)=2,808–8,487 (per country), 197,387 (total).

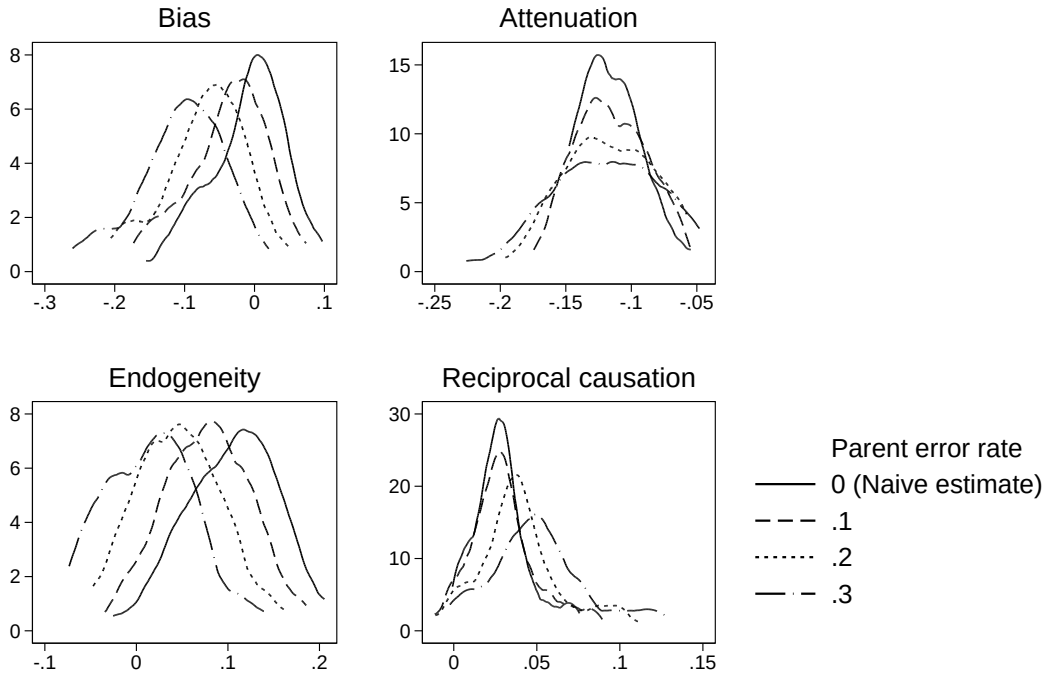

Figure 4: Simulation extrapolation estimates of bias components from Table 2 allowing for error in parent reports: 10%, 20%, and 30% misclassified. For further details on the assumed error structure, see running text and footnote 11. The first three panels correspond to the terms of equation (4), the last panel (bottom, right) to the middle term of equation (5). N=2,808–8,487 (per country), 197,387 (total).

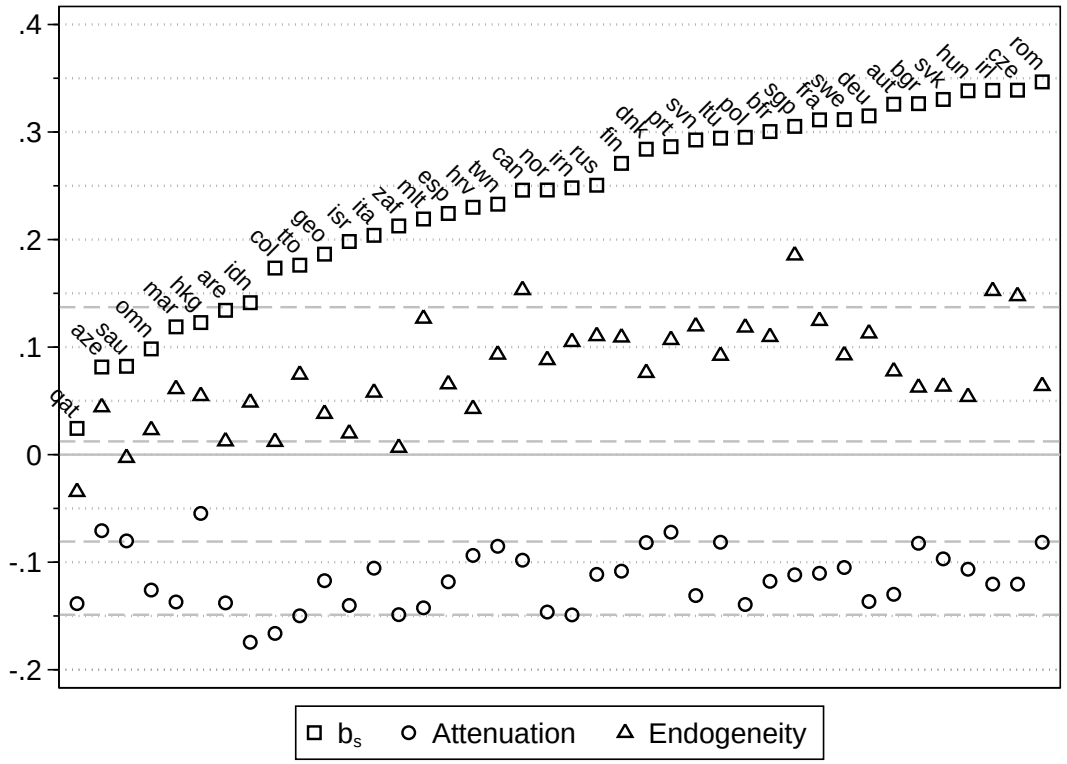

Figure 5: Regression coefficients of PIRLS 2011 reading score on student reported books at home (b_s), and estimated bias components based on validation against parent reports: attenuation and endogeneity, assuming 10% error in validation data. For country abbreviations, refer to Table 2. Dashed lines mark the 10 to 90 percentile range of each bias component. N=2,808–8,487 (per country), 197,387 (total).