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

Per Engzell

Nuffield College, University of Oxford
Swedish Institute for Social Research (SOFI), Stockholm University
per.engzell@nuffield.ox.ac.uk

Accepted for publication, *Sociological Methods & Research*
September 2018

Abstract

In studies of educational achievement, students' self-reported number of books in the family home is a frequently used proxy for social, cultural, and economic background. Absent hard evidence about what this variable captures or how well, its use has been motivated by strong associations with student outcomes. I show that these associations rest on two types of endogeneity: low achievers accrue fewer books, and are also prone to underestimate their number. The conclusion is substantiated both by comparing reports by students and their parents, and by the fact that girls report on average higher numbers despite being similar to boys on other measures of social background. The endogenous bias is large enough to overturn classical attenuation bias; it distorts cross-country patterns and invalidates many common study designs. These findings serve as a caution against overreliance on standard regression assumptions and contribute to ongoing debates about the empirical robustness of social science.

Keywords: *education, endogeneity, equality of opportunity, home literacy environment, socioeconomic status, standardized assessments, differential measurement error*

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 1916:100)

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). The aim of this study is to take a closer look at the measurement issues involved when the number of books in the home (henceforth, NBH) is used as an explanatory variable in models of student achievement. As of this article's writing, a search for "number of books [at/in the] home" returned close to 5,000 results in Google's *Scholar* database, two thirds of which were penned in the last decade.[1] Despite this popularity, surprisingly little is known about the measure's validity or reliability.

Instead, use of NBH has largely been motivated by one single consideration: its predictive power for student outcomes, of which already Holley wrote. For example, Hanushek and Woessmann (2011:117) recommend NBH as a proxy for students' social 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". Similarly, a 100-page methodological monograph issued by the International Association for the Evaluation of Educational Achievement (IEA) urged survey organizers to include those measures "that show the highest association with achievement in terms of explained variance", identifying NBH as "the strongest predictor of achievement . . . across the different studies and subject areas investigated" (Brese and Mirazchiyski 2013:98-99).

In this article I consider two explanations for the strong associations that have received little attention in previous studies. First, NBH may be an endogenous variable, if students who are good at reading garner more books as a result of their interests and abilities. Second, misreporting may not be random. Whereas in the classical measurement model any errors of observation in predictor variables will bias associations toward zero – the well-known "attenuation bias" – with systematic error the bias can go in any direction. In particular, students with little interest or ability in reading may underestimate NBH because they are unaware of any books that are available to them. This would contribute an upward bias, not downward as the classical measurement model posits. Inspecting a range of evidence, I find signs of both these types of endogeneity.

Using data from the PIRLS assessment of 10-year-olds in 40 countries, I start by establishing that student–parent agreement on this variable is low. I then provide evidence of systematic error: underreporting compared to parent responses is dominant and clearly related to low achievement. One possible objection to comparing the responses of students and parents is that the latter are likely to contain errors too. Therefore I corroborate my conclusions by inspecting gender differences as an exogenous source of achievement, unrelated to family background. This also allows me to assess the possibility that the actual number of books, and not just the student's error-prone estimate, is an endogenous variable. A later part of the paper goes on to illustrate the consequences of endogeneity in a cross-country comparative setting, using decomposition and simulation-based techniques adapted for the purpose. As this part of the analysis draws on stronger assumptions, it is worth noting that the evidence of endogeneity offered earlier in the paper is independent of it.

In conclusion, the high predictive power of NBH for achievement does not signal greater reliability or substantive importance compared to other proxies, and, ultimately, researchers may be better served by proxies that show more modest associations but are better measured. These results should challenge researchers to think more carefully about the assumptions that go into estimates based on proxy variables: while the classical model remains a convenient heuristic, its assumptions are not axiomatic and need to be justified in any given application. This study also adds to ongoing debates about limits to the self-correcting nature of social science (Gelman and

Loken 2014, Ioannidis 2012). While much of this debate has focused on sampling variability and selective reporting, the results uncovered here highlight the importance of endogeneity and mismeasurement as additional sources of bias. An implication is that the recent push toward increased transparency and replication (Freese and Peterson 2017), while laudable, may not be enough to rid social science of its biases without corresponding improvements in measurement and study design.

Background and Previous Literature

Measurement of students' social background is crucial in research on educational inequality and international differences in achievement (Breen and Jonsson 2005, van de Werfhorst and Mijs 2010). Virtually without exception, methodological literature on proxy variables departs from some version of the "classical" model where error is treated as random noise, unrelated to all model variables and the regression residual (Bohrnstedt 2010, Saylor 2013). It is well known that such error will lead to a bias toward the null as a simple function of the ratio of noise to total variance. This heuristic explains why researchers would take strong predictive power as an assurance for validity or reliability: if anything that can go wrong will lead to downward bias, proxies that predict the outcome well do so because they track the attributes proxied for more closely, are more reliably reported, or both.

However, the classical assumptions are just that – assumptions. As such they need to be substantiated, and large associations should never themselves be considered enough to validate a measure. As this study and a number of others show (Jerrim and Micklewright 2014, Rutkowski and Rutkowski 2010), reliability of NBH is not better than for other measures but in fact substantially worse. Meanwhile, there is little to suggest that NBH is strongly correlated with other observable aspects of students' social background. This raises the question: if superior reliability or validity do not explain the predictive power of NBH, then what does? This question is important given that many authors seem ready to accept NBH as a valid measure of family background based solely on its predictive power for student outcomes. Some have also invoked classical assumptions to correct for errors, to dramatic effect (Ammermueller and Pischke 2009), further motivating the question of whether such assumptions are indeed justified.

The question of *what NBH is a proxy for* receives somewhat different answers from study to study. Some use it to capture socioeconomic status broadly conceived – among them Hanushek and Woessmann (2011:117) who deem NBH "a powerful proxy for [students'] educational, social, and economic background". Others take it to reflect cultural as opposed to economic advantage (Esping-Andersen 2009, Evans, Kelley, and Sikora 2014, Marks, Cresswell, and Ainley 2006, Park 2008). Thus, Esping-Andersen (2009:128) reports that "'cultural capital' overpowers socioeconomic status in accounting for cognitive differences in all countries".[2] The aim here is not to adjudicate between these interpretations, but to assess whether methodological artefacts play a part in explaining the variable's strong predictive power.

In cross-country comparisons, the particular types of bias may matter less than whether they are similar across countries. Speaking to this issue, Schütz, Ursprung, and Woessmann (2008) regressed a banded measure of annual household income on NBH using parent-reported data from 6 countries in the Progress in International Reading Literacy Study (PIRLS). 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). 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.

This article builds on a string of recent studies that examined social background measurement in international student assessments (Engzell and Jonsson 2015, Jerrim and Micklewright 2014, Kreuter et al. 2010). Most closely related of these is Jerrim and Micklewright (2014) who studied NBH using the same dataset as here. They document not only low agreement between students and parents, but also wide variation in the strength of the association with student achievement depending on which source was used, with student reports often yielding the larger estimate. No previous study, however, has attempted to reconcile the low reliability with NBH's strong predictive power for student outcomes – the main contribution of this article. Methodologically, I also extend on recent discussions of systematic measurement error (Jerrim

and Micklewright 2014, Kreuter et al. 2010) by providing an empirical decomposition that separates classical attenuation from endogeneity and systematic misreporting as sources of bias. To this end, I adapt a method first used by Black, Sanders, and Taylor (2003), which is extended here to allow for imperfect validation data.

Data

Data for this study come from the Progress in International Reading Literacy Study (`https://timssandpirls.bc.edu`), conducted by the International Association for the Evaluation of Educational Achievement (IEA). PIRLS has collected information on NBH from students and parents 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.

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. 2009:11). 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 imputed as posterior draws from estimated ability distributions using a Rasch model (so-called "plausible values"). I standardize these to have mean=0, s.d.=1 within each country. All estimates account for uncertainty from plausible value imputation and clustering on school classes.[3]

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". 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 Programme for International

Student Assessment (PISA), asks students but not parents about books. I use PISA 2012 data in addition to PIRLS in the gender analyses below, where parent reports are not necessary. The PISA question is similar in its wording, but includes an additional category for "More than 500 books".

Descriptive Results

Previous studies have found that students and parents differ in their reports about NBH (Jerrim and Micklewright 2014, Rutkowski and Rutkowski 2010), but before turning to questions of the form and consequences of error, I revisit this issue briefly for two reasons. The questions asked in PIRLS are not identical, which could explain some of the discrepancy. Another possible explanation is age: it is well known that younger children are generally less reliable as respondents (Looker 1989). PIRLS participants are five years younger than those in PISA from whom comparable reliability estimates on occupation and education are available (Jerrim and Micklewright 2014).

Low Agreement between Students and Parents

Figure 1, left panel, plots Cohen's κ (kappa), a common measure of interreporter agreement, for student and parent-reported NBH in each PIRLS country. The statistics are trailing well below the 0.40 threshold commonly taken to denote "moderate" agreement (Landis and Koch 1977). As Table 1 shows, the questionnaire items differ: parents are asked about "books" and "children's books" separately, while students are asked for a total number. To approximate a comparable number with the parent questionnaire, I also construct an aggregate of both items 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.

The young age of students in PIRLS could be another issue. To address this, Figure 1 gathers comparable estimates of parent–child agreement from children close to PIRLS age (10 years). Although some are from small or nonrepresentative samples, they demonstrate that

higher agreement on other measures is not confined to the older PISA respondents.[4] Finally, Figure 1, right panel, displays rank order correlations. This is a more appropriate metric for the children's books item where the categories use different cutoffs, 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.[5] The upshot is that low agreement on NBH cannot be accounted for by questionnaire differences or student age.

The Structure of Disagreement

The κ statistics around 0.20 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 the resulting bias. As discussed above, if underreporting is more common among low achievers, the importance of books may be overstated in regression analyses that use student reports.

For now, I ignore the possibility of misreporting or endogeneity in parent reports and calculate the error in student reports as the difference relative to the total from the parent questionnaire. It is reasonable 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.

Using pooled data from all countries, 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 (0.46 vs. 0.15).

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 national achievement distribution increases the probability of an underreport by a factor of 1.6 (0.57 vs. 0.35). This difference is even starker in the category below ("11-25 books") with a factor of nearly two (0.41 vs. 0.22). Taking into account the extent of disagreement – the number of categories by which reports differ – accentuates these patterns even further (results not shown). This offers preliminary evidence that systematic misreporting makes student-reported data endogenous.

Learning from Gender Differences

The above findings are suggestive but may be sensitive to the assumption that parents report correctly. For this reason I turn to gender as an exogenous source of reading achievement. The intuition behind the gender comparison is simple. Boys and girls on average come from similar homes but girls outperform boys in reading throughout the school age (Buchmann, DiPrete, and McDaniel 2008).[6] If there is endogeneity, therefore, we would expect girls to report higher NBH. Because this strategy does not rely on linking sources I am also able to examine PISA data, where parents are not asked about NBH.

In addition to providing an independent test for endogeneity, gender differences can shed some light on its sources: it was noted above that NBH may be endogenous either because low achieving students (a) amass fewer books, (b) underreport, or (c) both. Endogenous underreporting would bring about a gender difference in student but not parent reports, while endogeneity in actual books would lead to a gender difference in both. If (a) is the case, therefore, we would expect a gender difference in NBH of similar size in both students and parents; if (b), a gender difference in students but not parents; and if (c), a gender difference in both, but of a larger magnitude in student reports which are subject to not one but two sources of endogeneity.

The results, reported as odds ratios from a set of ordered logistic regressions in Figure 3, confirm that girls tend to report greater numbers of books in both PISA and PIRLS. In the median country, a girl's odds of reporting in a higher category is 1.16 (PISA) or 1.15 (PIRLS)

times those of a boy. Figure 3 also reveals a similar, if smaller, differential by student gender in parent reports about "children's books". (While nominally the odds ratios are similar, this item spans a narrower range than the others, the maximum category being "More than 100"; see Table 1.) These results are further suggestive of endogeneity and on balance most consistent with scenario (c) where both reverse causation and endogenous underreporting contribute an upward bias.[7]

Consequences for Cross-Country Comparisons

The results so far are strongly suggestive of endogeneity and should lead to significant caution about NBH in any setting that attempts to estimate influences on student achievement. However, there are two questions that the above analysis does not answer. The first is whether endogeneity is quantitatively important; the second whether it distorts *cross-country patterns* in the association, on which much of this literature focuses (Chiu 2010, Jerrim and Micklewright 2012, Park 2008, Schütz et al. 2008). If endogeneity contributes a trivial bias compared to the standard problem of attenuation bias, concerns may be overwrought. Likewise if these issues manifest themselves similarly across countries, in which case the comparative picture would remain unchanged. To address these questions, I depart from a least-squares decomposition first used by Black, Sanders, and Taylor (2003) in a study of wage regressions. While Black et al. (2003) were primarily concerned with non-classical measurement error, I show here that their framework is useful for thinking about the wider problem of endogeneity.

To avoid thorny questions of what NBH is a proxy for I will simply assume that there is a true amount of books, which is what we ideally would like to observe. To make the assumption of exogeneity tenable, however, we should think of this as the number of books before the student came of reading age, or (more pedantically) the expected number of books at the time of survey, given parents' permanent characteristics. This obviously ignores a wide range of unobserved confounders, so exogeneity here should not be understood in the sense of identifying a *causal* effect of books, but only in the weaker sense that observed values are not themselves caused by student achievement. The books question is usually categorical ("0–10 books", "11–25 books", etc), but often modeled as linear in categories, assigning integer values such as 1 through 5 (e.g.,

Ammermueller and Pischke 2009, Esping-Andersen 2009, Jerrim and Micklewright 2012, Park 2008, Schütz et al. 2008). I follow this practice and to abstract from errors due to truncation or discretization, I further assume that the categories and not the underlying continuous variable are the target.

The inspection of gender differences above 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. Assuming that this variable reflects "the truth" might be going too far, however. Therefore I simulate the consequences of error in the validation data, to be interpreted as a sensitivity analysis in the spirit of Rosenbaum and Rubin (1983). I achieve this using a version of the simulation–extrapolation or *simex* algorithm (Cook and Stefanski 1994) described in greater detail below. I will also use the fact that parents report on "books" and "children's books" separately to assess the relative weight of reverse causation and systematic underreporting as sources of endogeneity.

Decomposition Method

The classical model on which nearly all work on proxy variables draws assumes that observed values x are an additive function of true values x^* and noise u, such that $x = x^* + u$. Given the classical assumptions that the error is mean zero, unrelated to true values, and to the regression residual, the least squares estimator regressing some outcome y on x is biased downward by a factor of noise to total variance (e.g., Bohrnstedt 2010):

$$plim\ \widehat{\beta}_{OLS} = \beta - \beta \frac{Var(u)}{Var(x)}$$

In our case, we are dealing with two types of deviations from the "true" underlying quantity: any books that have been brought into the house as a function of the student's reading achievement, and a response error reflecting the fact that the student may be misinformed, misread the question, or otherwise state the wrong answer. As argued above, neither of these deviations is likely to conform to classical assumptions. Imposing no assumptions on the form of u other

than additivity, the bias instead becomes:

$$plim\ \widehat{\beta}_{OLS} = \beta - \beta \underbrace{\frac{Cov(u,x)}{Var(x)}}_{\text{attenuation}} + \underbrace{\frac{Cov(u,\varepsilon)}{Var(x)}}_{\text{endogeneity}}$$

For proof and extended discussion of this and subsequent results, refer to the appendix at the end of this article. In fact, each of these components can be written as the slope coefficient from a separate regression, leading to the decomposition:

$$plim\ \widehat{\beta}_{OLS} = \beta - \beta(\beta_{ux}) + \beta_{\varepsilon x},$$

where β_{ux} is the slope from a regression of the error u on endogenous values x, and $\beta_{\varepsilon x}$ from regressing the residual ε on x (Black et al. 2003).

Of these components, the attenuation component β_{ux} is always positive, and when multiplied with $-\beta$ leads to a downward bias, just like in the classical model. It differs subtly from the classical expression, however, in that it also incorporates any correlation between the error and true values, x^*. Whenever a variable is bounded, floor and ceiling effects will entail that this correlation is negative, and attenuation is weakened compared to the classical case (Kreuter et al. 2010). The endogeneity component $\beta_{\varepsilon x}$ does not have a sign a priori, but depends on whatever process is generating the error. In our case, we know enough to say that this bias is upward in sign: it includes both endogeneity in the conventional sense (avid readers get more books) and the correlation between reporting error and achievement – which, if poor readers underestimate NBH, also biases the association upward.

Given validation data on correct as well as endogenous values, it is straightforward to estimate each of these components as:

- $\widehat{\beta}_{OLS}$: The slope from a regression of y on the endogenous measure x
- β: The slope from a regression of y on validation data x^*
- β_{ux}: The slope from a regression of the deviation $u = x - x^*$ on the endogenous measure x
- $\beta_{\varepsilon x}$: The slope from a regression of the residual $\varepsilon = y - \beta x^*$ on the endogenous measure x

As mentioned above, the validation data here come from parent reports about non-children's books, as the only exogenous measure available (in the weak sense of exogeneity above). While endogeneity renders parent-reported children's books unsuitable for this purpose, we can nevertheless use them to assess the relative sources of endogeneity: reciprocal causation versus systematic misreporting. The key assumption will be that by conditioning on parent reports about children's books, the remaining endogeneity is due to student misreporting. (This is a strong assumption, but recall that I also simulate the robustness of conclusions to the presence of error in parent reports.) Here, I simply reestimate the last component $\beta_{\varepsilon x}$ while flexibly controlling for parent-reported children's books, included as a set of indicator variables. Letting $\widetilde{\beta}_{\varepsilon x}$ denote the coefficient with children's books partialled out, we have:

$$\underbrace{\beta_{\varepsilon x}}_{\text{endogeneity}} = \underbrace{\beta_{\varepsilon x} - \widetilde{\beta}_{\varepsilon x}}_{\text{reciprocal causation}} + \underbrace{\widetilde{\beta}_{\varepsilon x}}_{\text{misreporting}}$$

Lastly, I simulate the consequence of error in parent reports using the *simex* algorithm introduced by Cook and Stefanski (1994) and adapted for categorical data by Küchenhoff, Mwalili, and Lesaffre (2006). Three scenarios are assessed, letting 10%, 20%, and 30% of parents misreport. While the details of this estimator are somewhat technically involved, the intuition is simple. It begins from the insight that while we cannot directly estimate the coefficient we would with perfect data, we can successively add more error and trace how the parameter of interest decays. Having done so, it is possible to fit a curve to the parameter decay and extrapolate back to the ideal case of no error. I apply the same amount of error to parent reports about "books" and "children's books" in these analyses. Further details of the simulation are described in the technical appendix.

Decomposition Results

I first focus on the limiting case of no error in the validation data; Table 2 shows point estimates from this decomposition, ordering countries by aggregate book possession ("Median" refers to the median category reported by parents). The regression estimates ("Student est.", "Parent est.") are generally in the range of 0.10 to 0.35 of a standard deviation's gain in reading scores per step up the "ladder" of categories. Given that the focus is on how cross-country *patterns* are

biased, I will have little to say about the substantive size of estimates – but the standard deviation of NBH averages about 1.3 (parents) or 1.2 (children) categories, so standardized coefficients are on the same order of magnitude.

The main interest lies with the subsequent columns, where "Bias" refers to the difference between regression estimates based on the two sources. This bias ranges from a negative –0.155 in Quatar, all but eradicating the parent-based estimate of 0.180, to a positive bias in countries such as Canada (0.075) or Singapore (0.097), as well as most European Union countries. Comparing countries with high and low aggregate numbers of books reveals a clear pattern. For countries where fewer books are the norm, toward the bottom of the table, parent reports yield larger estimates in line with classical measurement error. 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.

The next two columns ("Atten.", "Endog.") decompose the bias into attenuation and endogeneity. The attenuation component states just how much smaller the student-based estimate would be compared to that from parents, under the hypothetical scenario that the difference between the two reports was just random noise. In many cases, attenuation and endogeneity balance each other out so that the net bias comes close to zero. Nevertheless, that about half of the student estimate is then attributable to endogeneity – e.g., 48% in Norway or Germany (0.118/0.246; 0.150/0.315), or 50% in France (0.154/0.311) – means that when used in multivariate analyses, these different sources might yield quite different conclusions.

Endogeneity also tends to be more variable across countries than attenuation: the standard deviation of these two statistics across countries is 0.023 and 0.050, respectively. In other words, if attenuation was the only source of bias, the impact of family background would be understated but about equally much so in all countries. As it stands, the bias varies substantially across countries, largely as a function of variability in endogeneity. This point is depicted graphically in Figure 4, where the two components are shown alongside the student-based regression estimate, and countries ordered by the latter. Reading the plot from left to right,

the endogenous component grows in size with the substantive estimate – so that cross-country patterns estimated on student reports are attributable in no small part to differing endogeneity.

The next question is what happens once we relax the assumption of perfect validation data; results are found in Figure 5. The solid curve plots the distribution of estimates found in Table 2, assuming no error among parents. Dashed lines show how each set of estimates changes once we allow for the possibility of increased parental error: 10%, 20%, and 30% misclassified. The most obvious consequence is that β increases, which shifts the total bias downwards (top, left). At the same time, the extent of attenuation (top, right) is not much affected due to a simultaneous *decrease* in the the variance of the error, u. Instead, what explains the shift in total bias is a decline in the estimated endogeneity component (bottom, left). While this shows that the net sign of the bias may differ depending on the trust we are willing to put in parent-reported data, endogenous bias remains important.

A notable result appears in the last panel of Figure 5, which shows that the importance of reciprocal causation across specifications grows as that of endogenous misreporting fades. In Table 2 (last two columns), reciprocal causation accounted only for a negligible part of the endogenous bias, and that remains true with a modest parental misclassification rate of 10%. When we allow that rate to reach 20%, on the other hand, it contributes on average about half of the endogenous component, 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 2009:322). 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.

As a final check on the plausibility of decomposition results, it is worth comparing the cross-country pattern in endogeneity to the gender differentials in reporting mentioned earlier. This comparison is complicated by the fact that gender differences in NBH depends not only on

the extent of endogeneity, but also on the size of the boy–girl gap in reading. For comparability, I divide the gender difference in NBH with that in reading scores, for the 34 countries where the latter is statistically significant (p<.05). This yields a cross-country correlation of r=0.59 across the two sets of estimates. That figure rises to r=0.70 excluding countries where a majority of students attend single-sex schools, which could bias recruitment into the sample (Iran, Qatar, Saudi Arabia, United Arab Emirates). All in all, the pattern of endogenous bias across countries remains similar whether we take decomposition results or gender differences in reporting as a guide.

Conclusion

As a proxy for student background, self-reported books in the home are subject to endogeneity and systematic errors of observation. Not only do students from bookish homes perform better, but better students also accrue more books and are more informed about their home libraries. The resulting bias is large enough to outweigh the familiar attenuation bias, and lead to regression estimates of a similar size to those using parents as respondents. Guided by classical measurement theory, it is easy to misread the size of these estimates as signalling reliability or validity – with potentially damaging consequences for conclusions in the field.

Most obvious of these is perhaps that speculating about the influence of books or "culture" relative to other aspects of social background – measured by, for example, parents' education, social class, or economic status – will tell us little about actual transmission mechanisms; the dice will inevitably be loaded in favor of the endogenous measure, NBH. In cross-country comparisons, the endogenous bias appears to have about twice the variability as attenuation bias – ranging from being negligible, to accounting for as much as half or two thirds of estimated associations. While the specific figures change once we allow for imperfect validation data, the general conclusion remains and is corroborated by gender differences in reported books.

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 gradients vary by student age, student gender, or achievement domain such as reading or mathematics. NBH is ill suited as a proxy in each of these cases, as it is likely that endogeneity differs along several or all of these dimensions.

The problems uncovered here are likely to be exacerbated when attempts are made to correct for bias relying 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 with separate reports by two different individuals. As a consequence, they see their estimates triple in size. 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 negatively related to true values, meaning that an upward bias will result regardless (Kane, Rouse, and Staiger 1999).

It is notable that NBH has gained such widespread use, despite caution being voiced over a century ago (Holley 1916) and given that its flaws were hidden in plain sight. Arguably, this illustrates some of the forces that allow exaggerated results to proliferate in published literature – foremost, an academic culture that rewards storytelling and analyses that "work" over accuracy and robustness. In recent years, such practices have come under increased scrutiny (Gelman and Loken 2014, Ioannidis 2012). Yet, issues of measurement and misspecification have been conspicuously absent from these conversations, which have mostly revolved around a different set of concerns: underpowered research designs, flexibility in data collection and analysis, misuse of statistical tests, hypothesizing after results are known, and so on (Bernardi, Chakhaia, and Leopold 2017, Franco, Malhotra, and Simonovits 2014, Silberzahn et al. 2017, Simmons, Nelson, and Simonsohn 2011).

One upshot of all this is that recent initiatives for improved standards of transparency and replication (Freese and Peterson 2017, Muñoz and Young 2018), while important, may not be enough to rid social science of its biases. Such measures are designed to address selective reporting of "fluke" findings that result from sampling variability or arbitrary specification choices; they do not deal with systematic biases due to endogeneity or mismeasurement. It is

too early to tell whether sociology will suffer a replication crisis like that which has swamped some other disciplines in recent years (Open Science Collaboration 2015). But to the extent that it does not, sociologists should not be too ready to congratulate themselves. As this article illustrates, there is good reason to expect that a sanguine attitude to measurement and modeling may well be equally or more important as a source of spurious results in our discipline.

Notes

[1]These writings span the social sciences including sociology, psychology, economics, and education. The claim that NBH is the "single most important" predictor of achievement is a recurrent one (Ammermueller and Pischke 2009, Hanushek and Woessmann 2011, Schütz, Ursprung, and Woessmann 2008). A selective bibliography includes Brunello and Checchi (2007), Brunello, Weber, and Weiss (2017), Caro and Lenkeit (2012), Checchi and van de Werfhorst (2017), Chiu (2007, 2010), Chudgar and Luschei (2009), Esping-Andersen (2009), Ferreira and Gignoux (2014), Evans, Kelley, and Sikora (2014), Evans et al. (2010, 2015), Freeman and Viarengo (2014), Lurdes and Veiga (2010), Marks (2005), Marks, Cresswell, and Ainley (2006), Martins and Veiga (2010), Park (2008), Thorndike (1973), and Xu and Hampden-Thompson (2012). In economics, evidence on cross-country differences, peer effects, and 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). Findings have also been reported in popular media, influencing public discourse (e.g., *The New York Times*, 2011, 2015a,b).

[2]Some of these authors also veer toward a *causal* interpretation of the effect of books, Esping-Andersen (2008:128) stating that "children from a family with less than 10 books would enjoy a 9% improvement in their reading comprehension if parents were to arrive at the national average" of NBH, and Evans et al. (2014:13) claiming that "books matter enough to be policy relevant, with the gain from a 500-book home library equivalent to an additional three-quarters of a year of schooling".

[3]To economize on precision, survey weights are not applied (cf. Bollen et al. 2016), but results in Stapleton and Kang (2018) suggest that this choice does not make a large difference.

[4]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 0.57–0.72 in West et al. (N=1267–1476), 0.58–0.76 in Vereecken and Vandegehuchte (N=200), 0.43–0.63 in Ensminger et al. (N=119), and 0.34–0.63 in Andersen et al. (N=915).

[5]Engzell and Jonsson (2015:325) report Spearman's ρ from 14 year olds in the range of 0.41–0.59 for parental education and 0.62–0.74 for occupation (0.32–0.66 and 0.48–70 if the

parent was foreign born). Cohen and Orum (1972) report γ correlations from 9–13 year olds of 0.62.–72 for education and 0.75–0.85 for occupation. Andersen et al. (2008) report γ of 0.53–0.80 on their family affluence scale.

[6]Mullis et al. (2012:52) study these differences in PIRLS 2011 and report a female advantage of on average 16 score points, or roughly 1/6 of a standard deviation, but with marked variation across countries. Among the older children in PISA, the gender differential appears to be even larger (Salvi del Pero and Bytchkova 2013:22-23).

[7]Another notable pattern in Figure 3 is that girls report lower levels of parental education. As previous studies have found that exaggeration is the most common error for this variable (Kerckhoff, Mason, and Poss 1973), this is consistent with the idea that girls are more cognitively mature and more reliable as respondents in general (cf. Kreuter et al. 2010:131).

References

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.

Bernardi, Fabrizio, Lela Chakhaia, and Liliya Leopold. 2017. "Sing Me a Song with Social Significance: The (Mis)Use of Statistical Significance Testing in European Sociological Research." European Sociological Review, 33(1):1-15.

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.

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.

Bohrnstedt, George W. 2010. "Measurement Models for Survey Research." In: Rossi, P. H., Wright, J. D., Anderson, A. B., eds.: Handbook of Survey Research, 2nd edn. Academic Press.

Bollen, Kenneth A., Paul P. Biemer, Alan F. Karr, Stephen Tueller, and Marcus E. Berzofsky. 2016. "Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis." Annual Review of Statistics and Its Application, 3:375-392.

Borgers, Natacha, Edith De Leeuw, and Joop Hox. 2000. "Children as Respondents in Survey Research: Cognitive Development and Response Quality." Bulletin de methodologie sociologique, 66(1):60-75.

Breen, Richard, and Jan O. Jonsson. 2005. "Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility." Annual Review of Sociology, 31:223-243.

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. 2017. "Books Are Forever: Early Life Conditions, Education and Lifetime Earnings in Europe." Economic Journal, 127(600), 271-296.

Buchmann, Claudia, Thomas A. DiPrete, and Anne McDaniel. 2008. "Gender Inequalities in Education." Annual Review of Sociology, 34:319-337.

Caro, Daniel H., and Jenny Lenkeit. 2012. "An Analytical Approach to Study Educational Inequalities: 10 Hypothesis Tests in PIRLS 2006." International Journal of Research & Method in Education 35(1):3-30.

Checchi, Daniele, and Herman G. van de Werfhorst. 2017. "Policies, Skills and Earnings: How Educational Inequality Affects Earnings Inequality." Socio-Economic Review 16(1):137-160.

Chiu, Ming Ming. 2007. "Families, Economies, Cultures, and Science Achievement in 41 Countries: Country-, School-, and Student-Level Analyses." Journal of Family Psychology, 21(3):510-519.

Chiu, Ming Ming. 2010. "Effects of Inequality, Family and School on Mathematics Achievement: Country and Student Differences." Social Forces, 88(4):1645-1676.

Chudgar, Amita, and Thomas F. Luschei. 2009. "National Income, Income Inequality, and the Importance of Schools: A Hierarchical Cross-National Comparison." American Educational Research Journal, 46(3):626-658.

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.

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.

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.

Esping-Andersen, Gøsta. 2009. The Incomplete Revolution. Cambridge: Polity.

Evans, M.D.R., Jonathan Kelley, and Joanna Sikora. 2014. "Scholarly Culture and Academic Performance in 42 Nations." Social Forces, 92(4):1573-1605.

Evans, M.D.R., Jonathan Kelley, Joanna Sikora, and Donald J. Treiman. 2010. "Family Scholarly Culture and Educational Success: Books and Schooling in 27 Nations." Research in Social Stratification and Mobility, 28(2):171-197.

Evans, M.D.R., Jonathan Kelley, Joanna Sikora, and Donald J. Treiman. 2015. "Scholarly Culture and Occupational Success in 31 Societies." Comparative Sociology, 14(2):176-218.

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

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." Science 345(6203):1502-05.

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

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." Annual Review of Sociology 43:147-165.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science: Data-Dependent Analysis – A 'Garden of Forking Paths' – Explains Why Many Statistically Significant Comparisons Don't Hold Up." American Scientist 102(6):460-465.

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.

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

Ioannidis, John. 2012. "Why Science Is Not Necessarily Self-Correcting." Perspectives on Psychological Science 7(6):645-654.

Jerrim, John, and John Micklewright. 2012. "Parental Socio-Economic Status and Children's Cognitive Achievement at Ages 9 and 15: How Do the Links Vary Across Countries?" In Ermisch, J., Jäntti, M., Smeeding, T. M., eds.: From Parents to Children: The Intergenerational Transmission of Advantage. New York: Russell Sage.

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.

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger. 1999. "Estimating Returns to Schooling When Schooling Is Misreported." Working Paper no. 7235, National Bureau of Economic Research.

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.

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.

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

Marks, Gary N. 2005. "Cross-National Differences and Accounting for Social Class Inequalities in Education." International Sociology, 20(4):483-505.

Marks, Gary N., John Cresswell, and John Ainley. 2006. "Explaining Socioeconomic Inequalities in Student Achievement: The Role of Home and School Factors." Educational Research and Evaluation, 12(2):105-128.

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.

Micklewright, John, and Sylke V. Schnepf. 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, Ann M. Kennedy, Kathleen L. Trong, and Marian Sainsbury. 2009. PIRLS 2011 Assessment Framework. Amsterdam: IEA.

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

Muñoz, John, and Cristobal Young. 2018. "We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness." Sociological Methodology, early access. doi: 10.1177/0081175018777988

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.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." Science, 349(6251):aac4716.

Park, Hyunjoon. 2008. "Home Literacy Environments and Children's Reading Performance: A Comparative Study of 25 Countries." Educational Research and Evaluation, 14(6):489-505.

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.

Salvi del Pero, Angelica, and Alexandra Bytchkova (2013). "A Bird's Eye View of Gender Differences in Education in OECD Countries", OECD Social, Employment and Migration Working Papers, No. 149, OECD Publishing. doi: 10.1787/5k40k706tmtb-en

Saylor, Ryan. 2013. "Concepts, Measures, and Measuring Well: An Alternative Outlook." Sociological Methods & Research, 42(3):354-391.

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

Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli C. Awtrey, Štěpán Bahník, et al. 2017. "Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results." PsyArXiv. September 21. doi: 10.31234/osf.io/qkwst

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." Psychological Science, 22(11):1359-1366.

Stapleton, Laura M., and Yoonjeong Kang. 2018. "Design Effects of Multilevel Estimates from National Probability Samples." Sociological Methods & Research, 47(3):430-457.

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. Berkeley, CA: McCutchan.

van de Werfhorst, Herman G., and Jonathan J. B. Mijs. 2010. "Achievement Inequality and the Institutional Structure of Educational Systems: A Comparative Perspective." Annual Review of Sociology, 36:407-428.

Vereecken, Carine, and Ann Vandegehuchte. 2003. "Measurement of Parental Occupation: Agreement between Parents and Their Children." Archives of Public Health, 61:141-149.

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.

Xu, Jun, and Gillian Hampden-Thompson. 2012. "Cultural Reproduction, Cultural Mobility, Cultural Resources, or Trivial Effect? A Comparative Approach to cultural capital and Educational Performance." Comparative Education Review, 56(1):98-124.

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	0.246 (0.018)	0.265 (0.019)	–0.019	–0.137	0.118	0.022	0.096
Sweden (swe)	3837	101–200	0.312 (0.014)	0.295 (0.014)	0.016	–0.111	0.128	0.025	0.103
Hungary (hun)	4832	26–100	0.338 (0.017)	0.354 (0.017)	–0.015	–0.117	0.101	0.031	0.071
Denmark (dnk)	4299	26–100	0.284 (0.015)	0.262 (0.014)	0.022	–0.089	0.111	0.016	0.094
Germany (deu)	2960	26–100	0.315 (0.017)	0.302 (0.016)	0.013	–0.137	0.150	0.028	0.122
Georgia (geo)	4416	26–100	0.186 (0.014)	0.243 (0.019)	–0.056	–0.118	0.061	0.011	0.050
Finland (fin)	4368	26–100	0.271 (0.016)	0.241 (0.014)	0.030	–0.108	0.138	0.032	0.105
Austria (aut)	4356	26–100	0.326 (0.015)	0.342 (0.013)	–0.016	–0.135	0.119	0.037	0.081
Czech Rep (cze)	4335	26–100	0.339 (0.016)	0.276 (0.015)	0.063	–0.120	0.183	0.019	0.163
Canada (can)	18471	26–100	0.246 (0.008)	0.171 (0.007)	0.075	–0.094	0.170	0.010	0.160
Ireland (irl)	4149	26–100	0.339 (0.014)	0.280 (0.013)	0.059	–0.124	0.183	0.036	0.147
Malta (mlt)	3154	26–100	0.219 (0.020)	0.215 (0.016)	0.004	–0.137	0.141	0.040	0.101
Spain (esp)	7827	26–100	0.224 (0.013)	0.251 (0.012)	–0.026	–0.118	0.091	0.025	0.067
Russia (rus)	4399	26–100	0.251 (0.020)	0.226 (0.019)	0.024	–0.109	0.134	0.027	0.107
Belgium Fr (bfr)	3300	26–100	0.300 (0.020)	0.285 (0.017)	0.016	–0.122	0.138	0.033	0.105
France (fra)	4019	26–100	0.311 (0.016)	0.273 (0.015)	0.038	–0.116	0.154	0.034	0.120
Slovakia (svk)	5414	26–100	0.330 (0.018)	0.327 (0.018)	0.003	–0.105	0.108	0.052	0.056
Israel (isr)	3213	26–100	0.198 (0.020)	0.291 (0.019)	–0.093	–0.141	0.048	0.067	–0.020
Bulgaria (bgr)	5041	26–100	0.326 (0.020)	0.321 (0.020)	0.005	–0.096	0.101	0.021	0.081
Poland (pol)	4843	26–100	0.295 (0.015)	0.282 (0.013)	0.013	–0.137	0.150	0.027	0.122
Italy (ita)	3806	26–100	0.204 (0.015)	0.231 (0.017)	–0.027	–0.107	0.080	0.028	0.052
Slovenia (svn)	4274	26–100	0.293 (0.016)	0.271 (0.013)	0.022	–0.129	0.151	0.048	0.103
Portugal (prt)	3845	26–100	0.286 (0.017)	0.231 (0.015)	0.055	–0.081	0.136	0.033	0.102
Lithuania (ltu)	4367	26–100	0.294 (0.019)	0.257 (0.015)	0.038	–0.087	0.125	0.040	0.084
Trinidad (tto)	3422	26–100	0.176 (0.020)	0.222 (0.019)	–0.046	–0.140	0.094	0.024	0.070
Taiwan (twn)	4192	26–100	0.233 (0.013)	0.206 (0.013)	0.027	–0.089	0.116	0.010	0.106
Singapore (sgp)	6077	26–100	0.305 (0.015)	0.208 (0.013)	0.097	–0.108	0.205	0.064	0.142
Croatia (hrv)	4457	26–100	0.230 (0.016)	0.250 (0.014)	–0.020	–0.095	0.075	0.028	0.047
Romania (rom)	4401	26–100	0.347 (0.021)	0.340 (0.018)	0.006	–0.097	0.103	–0.001	0.104
Hong Kong (hkg)	3487	26–100	0.123 (0.020)	0.112 (0.017)	0.011	–0.055	0.066	0.025	0.041
Qatar (qat)	3413	26–100	0.024 (0.016)	0.180 (0.020)	–0.155	–0.130	–0.025	0.030	–0.055
UA Emirates (are)	12709	11–25	0.134 (0.013)	0.237 (0.013)	–0.103	–0.135	0.032	0.076	–0.044
Saudi Arabia (sau)	4216	11–25	0.082 (0.022)	0.150 (0.021)	–0.068	–0.079	0.011	0.012	–0.000
Oman (omn)	8752	11–25	0.098 (0.012)	0.174 (0.011)	–0.076	–0.114	0.038	0.024	0.014
Azerbaijan (aze)	4272	11–25	0.081 (0.020)	0.091 (0.020)	–0.010	–0.062	0.052	0.005	0.047
South Africa (zaf)	2605	11–25	0.213 (0.036)	0.313 (0.031)	–0.100	–0.145	0.045	0.027	0.018
Iran (irn)	5515	11–25	0.248 (0.019)	0.255 (0.018)	–0.007	–0.140	0.133	0.047	0.086
Colombia (col)	3669	11–25	0.174 (0.032)	0.270 (0.030)	–0.097	–0.147	0.050	0.017	0.033
Morocco (mar)	5474	0–10	0.119 (0.024)	0.158 (0.023)	–0.039	–0.114	0.076	–0.001	0.076
Indonesia (idn)	4400	0–10	0.141 (0.042)	0.217 (0.032)	–0.075	–0.147	0.072	0.006	0.066

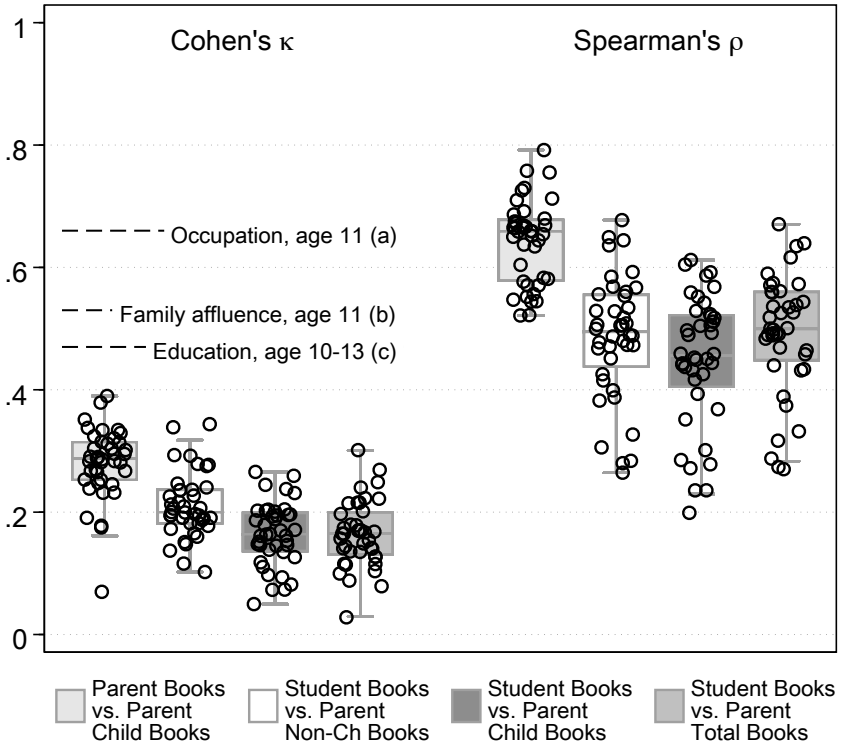

Figure 1: Agreement between students and parents on books in the home in PIRLS 2011. Cohen's κ (left) and Spearman's ρ (right). Each circle represents a country. The items are described in Table 1 and the running text. Median κ estimates from earlier studies are displayed for comparison (dashed lines), 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).

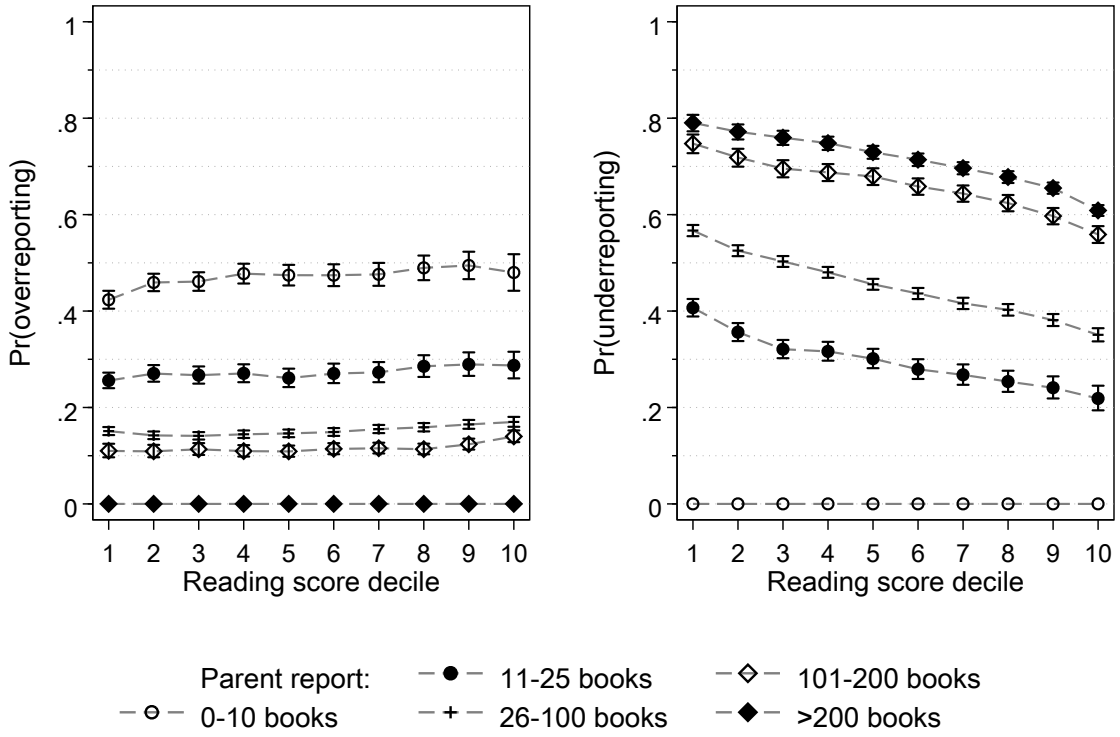

Figure 2: Estimated probabilities from fully interacted logistic regression of students reporting a higher or lower category than parent ("over" and "underreporting"), by student's achievement decile and parent's reported value. Pooled data from PIRLS 2011, achievement scores standardized at the country 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.

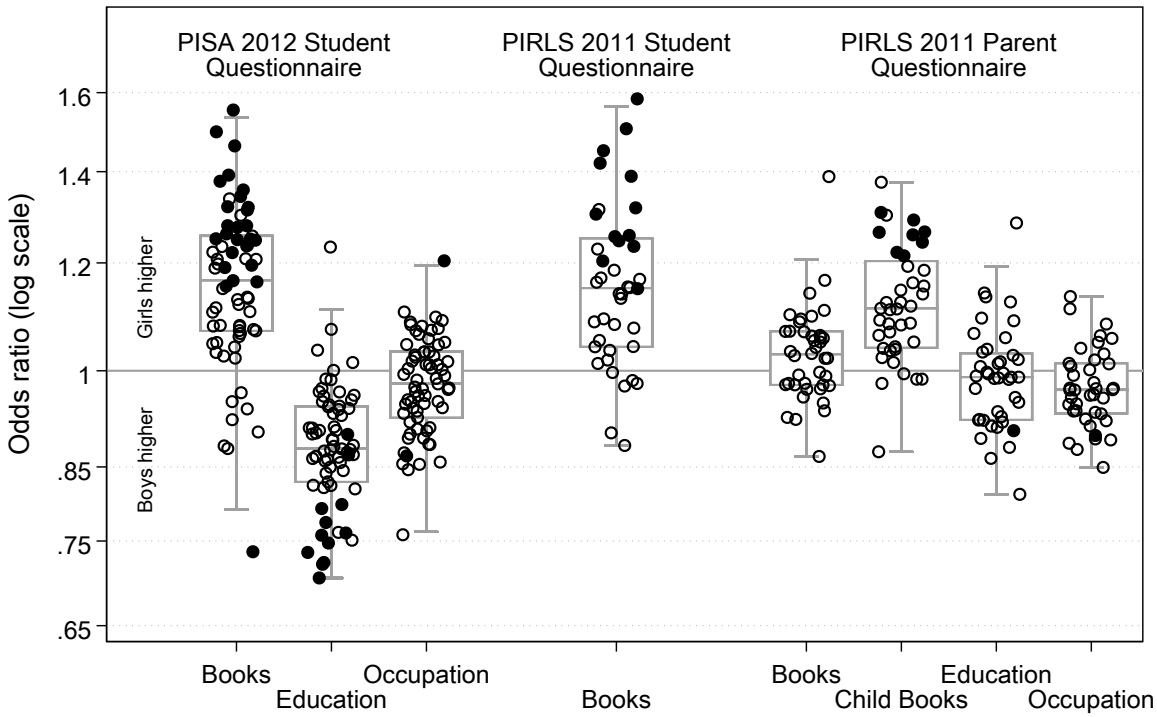

Figure 3: Odds ratios from ordered logistic regression of reported student background variables on student gender. Each circle represents a country. Filled markers indicate significance at the 5% confidence level, Bonferroni corrected by the number of study countries and allowing for clustering on the school (PISA) or school class (PIRLS) level. Box plots display the median and interquartile range of estimates. 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).

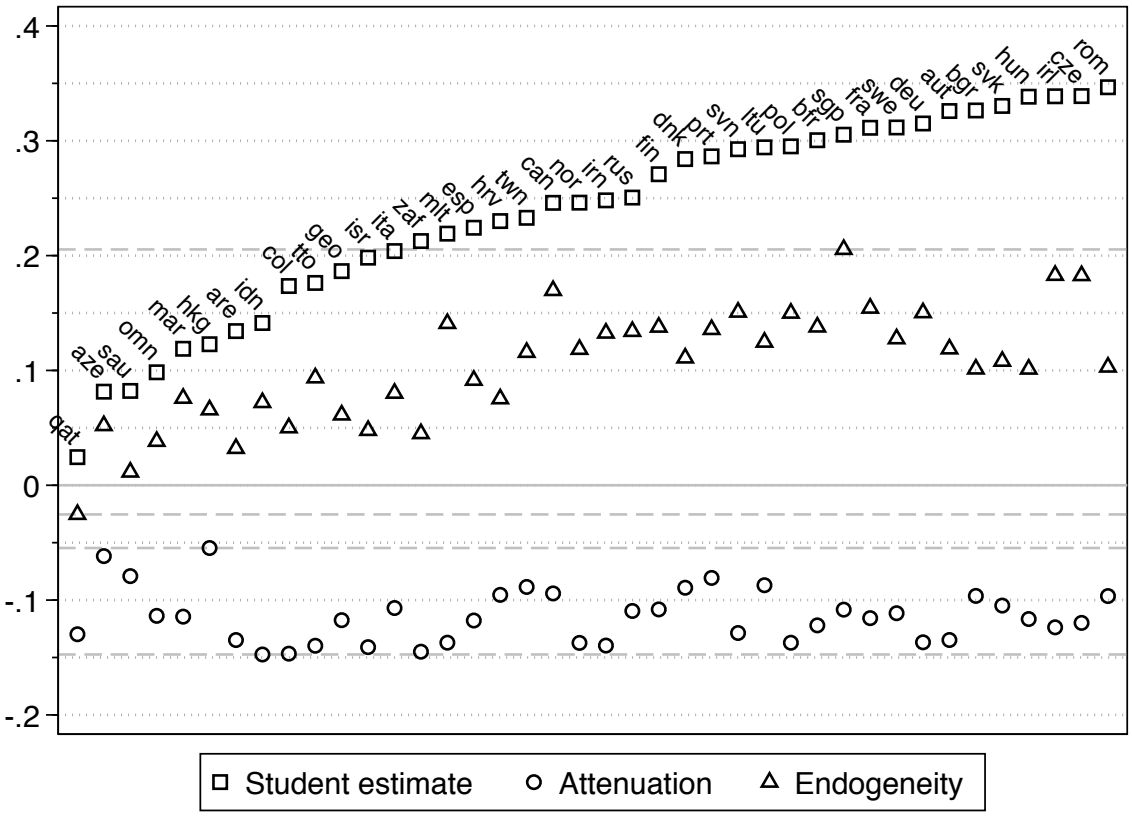

Figure 4: Regression coefficients of PIRLS 2011 reading score on student-reported books at home (b_s), and estimated bias components based on validation against parent reports: attenuation and endogeneity, assuming no error in validation data. Dashed lines mark the range of each bias component. All estimates are from Table 2, which also provides key for the country abbreviations. N=2,808–8,487 (per country), 197,387 (total).

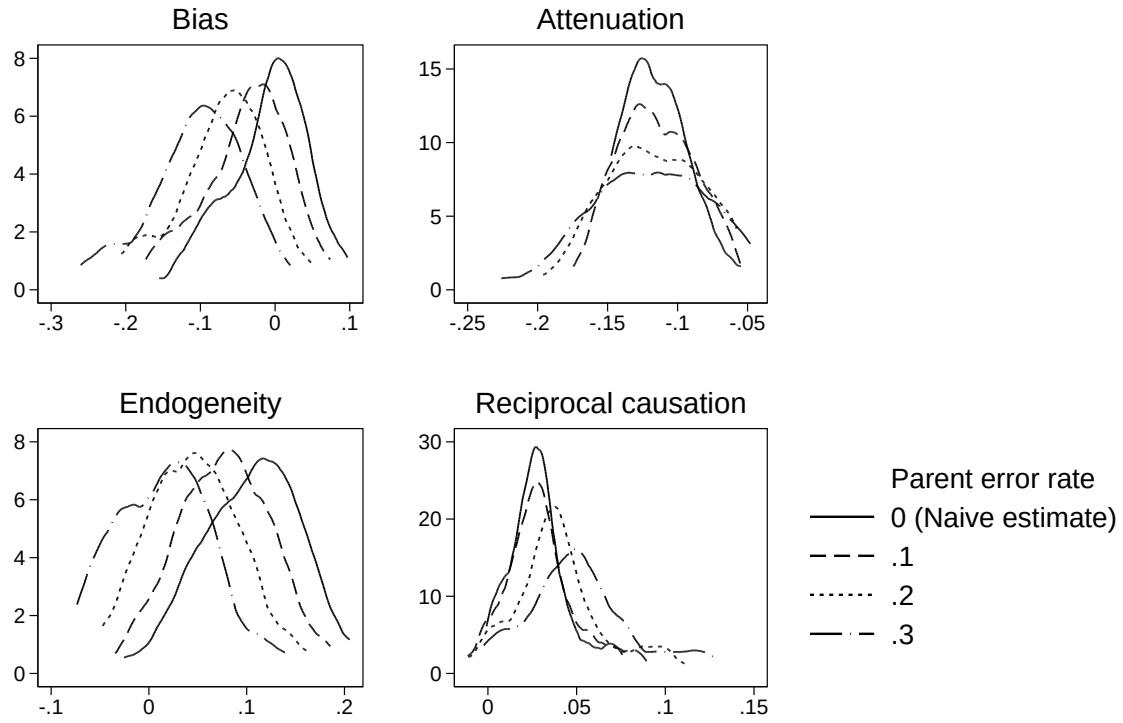

Figure 5: Simulation extrapolation estimates of bias components from Table 2 allowing for error in parent reports: 10%, 20%, and 30% misclassified. For further details on the assumed error structure, see running text and note 11. The first three panels correspond to the terms of equation (4), the last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

Technical Appendix

The expression for bias with a noisy and endogenous proxy is derived as follows. Assume that observed values x are the sum of true values x^* and a noise term u. Furthermore, write the population regression slope $\beta = Cov(x^*, y)/Var(x^*)$, and the residuals from this regression ε. The ordinary least squares estimator with the erroneous variable has probability limit:

$$\begin{aligned}
plim\ \widehat{\beta}_{OLS} &= \frac{Cov(x,y)}{Var(x)} \\
&= \frac{Cov(x,(\beta x - \beta u + \varepsilon))}{Var(x)} \\
&= \beta + \frac{Cov(x,(-\beta u + \varepsilon))}{Var(x)} \\
&= \beta - \beta\frac{Cov(x,(u+\varepsilon))}{Var(x)} \\
&= \beta - \beta\frac{Cov(x,u)}{Var(x)} + \frac{Cov(x,\varepsilon)}{Var(x)} \\
&= \beta - \beta\frac{Cov(u,x)}{Var(x)} + \frac{Cov(u,\varepsilon)}{Var(x)}
\end{aligned}$$

That $Cov(x,\varepsilon)$ and $Cov(u,\varepsilon)$ are interchangeable in the last step follows from assuming that $x = x^* + u$, together with $Cov(x^*,\varepsilon) = 0$ which is true by construction. In fact, the knowledge that $x = x^* + u$ allows us to go further and show where the classical model comes from:

$$\begin{aligned}
plim\ \widehat{\beta}_{OLS} &= \beta - \beta\frac{Cov(u,x)}{Var(x)} + \frac{Cov(u,\varepsilon)}{Var(x)} \\
&= \beta - \beta\frac{Cov(u,u) + Cov(u,x^*)}{Var(x)} + \frac{Cov(u,\varepsilon)}{Var(x)} \\
&= \beta - \beta\frac{Var(u)}{Var(x)} - \beta\frac{Cov(u,x^*)}{Var(x)} + \frac{Cov(u,\varepsilon)}{Var(x)}
\end{aligned}$$

Since $Cov(u,x^*)$ and $Cov(u,\varepsilon)$ are both assumed to be zero in the classical model, the last two terms drop out and we are left with the standard noise-to-total-variance ratio.

The main text outlines how each of these components can be estimated given a validation dataset; in our case, the gold standard is that of parents. To further assess robustness to errors in parent-reported data, I use the method of simulation–extrapolation (*simex*), which offers a general approach for handling measurement error that cannot be assumed normal in form (Cook and Stefanski 1994, Küchenhoff et al. 2006). In brief, the idea of this estimator is to:

(1) reestimate the model while repeatedly adding simulated noise of a specified form, (2) fit a regression curve to parameter decay as a function of the amount of added error, (3) extrapolate back to the ideal case of no error to infer what the estimate would have been, had no error been present.

As a Monte Carlo-based estimator, simex makes minimal parametric assumptions. In particular, it does not impose any distribution on the unobserved regressor 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 – unlike that in students – is exogenous, 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, letting 10%, 20%, and 30% of parents misreport.

The 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. Pseudo data are then 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 $\{0, 0.25, 0.5, \ldots 2\}$ 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.

To structure the off-diagonals of $\mathbf{\Pi}$, 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 random misclassification causes more rapid parameter decay. I also allow parents' errors to the two questions to be correlated ($r \approx .3$). Taking errors to be orthogonal, the role of reciprocal causation remains minor across all specifications but other results remain unchanged. Conversely, allowing maximally correlated errors increases the contribution of reciprocal causation. 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.