Rew *et al.*
Large-scale Assessments in Education (2022) 10:20
https://doi.org/10.1186/s40536-022-00139-3

Large-scale Assessments in Education

RESEARCH Open Access

Student achievement, school quality, and an error-prone family background measure: exploring the sensitivity of the Heyneman-Loxley effect in Southern and Eastern Africa

W. Joshua Rew[1*], Anabelle Andon[2] and Thomas F. Luschei[3]

*Correspondence:
josh.rew@ode.oregon.gov

[1] Oregon Department of Education, Salem, USA
[2] Columbia University, New York, USA
[3] Claremont Graduate University, Claremont, USA

Abstract

Background: We examine the sensitivity of the Heyneman-Loxley Effect to the influence of an error-prone family background measure in 15 education systems from Southern and Eastern Africa. Our aim is to revisit a claim by Abby Riddell from the November 1989 issue of the *Comparative Education Review* concerning the reliability of family background measures and the estimation of the Heyneman-Loxley Effect. Three questions guide our study: does national income have an association with the reliability of a family background measure, is the association between a family background measure and student achievement sensitive to measurement error, and is the association between national income and the school effect sensitive to measurement error?

Methods: Our analysis relies on the SACMEQ III data archive and, most importantly, a known error-prone family background measure (i.e., socioeconomic status index) and its corresponding measurement error (i.e., conditional standard error of measurement). For each SACMEQ III education system, we calculate the reliability of the socioeconomic status index and examine its association with national income. We use a Bayesian multilevel regression model to estimate naive and correction parameters representing the association between the socioeconomic status index and student achievement. Finally, we explore the associations between national income and the naive and correction estimates for the school effect across SACMEQ III education systems.

Results: We observe three results. First, the association between national income and the reliability of the socioeconomic status index appears negative among SACMEQ III education systems (albeit questionable due to the small n-size and influential outliers). Second, the association between the socioeconomic status index and student achievement is sensitive to measurement error across content areas and SACMEQ III education systems. Third and finally, the association between national income and the school effect is insensitive to measurement error across content areas and SACMEQ III education systems.

Conclusions: Throughout our study, we discuss measurement error, its consequences, and why the correction of error-prone family background measures is important. We highlight the need for auxiliary information for measurement error correction (e.g., reliability ratio, conditional standard error of measurement). Lastly, in addition to encouraging the correction of error-prone family background measures when attempting to replicate the Heyneman-Loxley Effect, we invite further research on improving the reliability and comparability of family background measures.

Keywords: Heyneman-Loxley Effect, Student achievement, Socioeconomic status, Measurement error, Bayesian estimation, Southern and Eastern Africa, SACMEQ

Introduction

Since the publication of *Equality of Educational Opportunity* in 1966 (Coleman et al., 1966),[1] national and cross-national studies have sought to question, confirm, or replicate the report's findings that family background explains more variation in student achievement than the quality of schooling (e.g., Atteberry & McEachin, 2020; Rury & Saatcioglu, 2015; Konstantopoulos & Borman, 2011; Borman & Dowling, 2010; Hanushek, 1997; Greenwald, Hedges, & Laine, 1996; Simmons & Alexander, 1978; Summers & Wolfe, 1977; Carver, 1975; Cain & Watts, 1970; Bowles & Levin, 1968). The implications of this "Coleman Effect," whether seen as controversial or remarkable, became the inspiration for studies from a diverse set of academic disciplines, fields, societies, and journals, including sociology, economics, comparative and international education, and school effectiveness and school improvement.[2]

A noteworthy rebuttal to the Coleman Effect came from a cross-national study by Heyneman and Loxley (1983). Using student achievement data from 29 countries, Heyneman and Loxley found that the association between the quality of schooling and student achievement was not uniform across countries. That is, in countries with lower national incomes, the quality of schooling had a greater influence on student achievement than family background. This phenomenon became known as the "Heyneman-Loxley (HL) Effect" (Baker et al., 2002) and, similar to the Coleman Effect, received considerable attention, debate, and challenge from national and cross-national studies (e.g., Baker et al., 2002; Bouhlila, 2015; Chiu, 2007; Chmielewski, 2019; Chudgar & Luschei, 2009; Gamoran & Long, 2007; Gruijters & Behrman, 2020; Hanushek & Luque, 2003; Harris, 2007; Huang, 2010; Ilie & Lietz, 2010; Lee & Borgonovi, 2022; Riddell, 1989a; Schiller et al., 2002; Woessmann, 2010).

A significant challenge to the HL Effect came from Riddell in the November 1989 issue of the *Comparative Education Review*. As part of a special focus on challenges to prevailing theories, Riddell (1989a) identified several methodological limitations to estimating the HL Effect. The most prominent limitations were the use of ordinary least squares (OLS) regression and r^2 (rather than multilevel regression modeling and the intraclass correlation coefficient [ICC]) to determine the importance of school

[1] Although *Equality of Educational Opportunity* continues to receive attention and recognition, it is necessary to acknowledge the publication and importance of *Children and their Primary Schools* chaired by Lady Bridget Plowden (Department of Education and Science, 1967).

[2] The mission statement by Reynolds and Creemers (1990) in the inaugural issue of *School Effectiveness and School Improvement* was a response to years of research favoring the Coleman Effect.

quality relative to family background in both lower income and wealthy countries.[3] However, a less publicized limitation implied by Riddell was the questionable reliability of family background measures in lower income countries. This arose as part of Riddell's brief discussion of r^2 and its susceptibility to misestimation due to the use of error-prone measures (e.g., family background). Riddell discreetly suggested that, if family background measures were less reliable in lower income countries than in wealthy countries, the HL Effect may simply be an artifact of measurement error.

The reliability of family background measures did not receive further attention in Heyneman's commentary (1989), Riddell's rejoinder (1989b), or in later research until Baker et al. (2002) attempted to replicate the HL Effect using data from the Third International Mathematics and Science Study (TIMSS). Upon finding that the association between the quality of schooling, family background, and student achievement was reasonably uniform across countries (reflecting the Coleman Effect rather than the HL Effect), Baker, Goesling, and LeTendre reconsidered the limitations originally identified by Riddell. While they confirmed that less reliable family background measures pose a threat to the estimation of the HL Effect, Baker, Goesling, and LeTendre did not find evidence to corroborate Riddell's claim. Later studies have largely overlooked the questionable reliability of family background measures as they attempted to substantiate or refute the HL Effect. An exception is the study by Chudgar and Luschei (2009). As part of their discussion of limitations, Chudgar and Luschei warned that error-prone covariates, especially family background measures, could bias the associations between student achievement, school quality, and family background (and inevitably lead to the misestimation of the HL Effect).

Given the claim by Riddell, the efforts of Baker et al. (2002), and the warnings from Chudgar and Luschei (2009), the aim of our study is to explore the sensitivity of the HL Effect to an error-prone family background measure. We rely on student achievement and family background data from the 15 education systems that participated in the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) III project. Unlike prior research, our analysis benefits from the availability of a known error-prone family background measure and its respective measurement error. With this additional information and using a Bayesian multilevel regression model, we calculate the reliability of the family background measure, and estimate both naive and correction parameter estimates for each education system and content area (i.e., reading, mathematics, and HIV/AIDS knowledge). The results from these analyses allow us to address the following research questions:

1. *Does national income have an association with the reliability of a family background measure?*
2. *Is the association between a family background measure and student achievement sensitive to measurement error?*

[3] In a commentary to Riddell, Heyneman (1989) stated that "the Riddell article implies some dismay about the good judgment of users of ordinary least squares (OLS) analytic techniques in the 1970s. But this is like faulting Charles Lindbergh for not using radar" (p. 498). Heyneman (2016) later added that critiquing the use of OLS regression instead of multilevel regression was akin to "attacking Charles Lindbergh for not using radar, a technology that had not yet been invented" (p. 155).

3. *Is the association between national income and the school effect sensitive to measurement error?*

Our study contains four sections. First, we define and discuss measurement error, the consequences of ignoring error-prone family background measures, and how to correct for measurement error (particularly the need for auxiliary information). Second, we describe our methods, which include the data sources we use, our analyses, and assumptions that support our analyses. Third, we summarize the results from our analyses, address the three research questions, and discuss their implications. Fourth and finally, we conclude by exploring the association between a family background measure and its respective measurement error, and inviting further research on improving the reliability and comparability of family background measures.

Measurement error

Measurement error is the difference between an observable manifestation (or the combination of manifestations) and the true but unobservable value of the respective latent trait (Buonaccorsi, 2010; Carroll et al., 2006; Fuller, 2009). In economics, psychology, sociology, and other fields within the social and behavioral sciences, it is quite difficult to directly observe certain phenomena (Fuller, 2009; Bound, Brown, & Mathiowetz, 2001). Unobservable phenomena are known as latent traits and examples include teacher attitudes and beliefs, student academic knowledge and skills, and household socioeconomic status (SES). Researchers often administer questionnaires to students, parents, educators, and school administrators to collect the observable manifestations of these traits. For instance, many international large-scale assessment programs (e.g., Progress in International Reading Literacy Study [PIRLS], Program for International Student Assessment [PISA], and TIMSS) collect observable manifestations of SES from students in target grades or age groups. Examples of these manifestations include the number of books in the home, specific household possessions, parent education, and parent occupational status (Avvisati, 2020; Buchmann, 2002; Chudgar et al., 2014; Yang & Gustafsson, 2004).

Observable manifestations are imprecise estimates of their respective latent traits. This imprecision is largely due to the self-report nature of questionnaires where responses, especially from students, are subject to guessing, estimation, and imperfect recall (Hannum et al., 2017; Ridolfo & Maitland, 2011; Fargas-Malet et al., 2010; Fowler Jr. & Cosenza, 2009). In the context of measuring SES, students may not know, remember, or wish to respond to items regarding the education level of their parents (Engzell & Jonsson, 2015; Kreuter, Eckman, Maaz, & Watermann, 2010; Bos & Kuiper, 1999), the number of books in their home (Engzell, 2021; Rutkowski & Rutkowski, 2018), or specific household possessions (Jerrim & Micklewright, 2014). For younger and low achieving students, this typically leads to nonresponse or guessing, and, ultimately, the misestimation of the association between student SES, achievement, and other student outcomes within and across countries (Chudgar et al., 2014; Engzell, 2021). Solutions taken by a few international large-scale assessment programs to minimize measurement error include omitting parent education items from lower grade questionnaires (e.g., TIMSS 4th grade student questionnaire), simultaneously administering a

questionnaire to parents or guardians (e.g., PIRLS learning to read survey), or sending the questionnaire home for parents to verify student responses (e.g., SACMEQ III homework form; Cresswell, Schwantner, & Waters, 2015).

Observable manifestations also reflect partial or incomplete information concerning the latent trait. Researchers often make decisions prior to or after the administration of questionnaires in order to decrease the response burden of students, protect student confidentiality, or lessen questionnaire administration, analysis, and reporting expenses (Yi & Buzas, 2021). These decisions may inadvertently decrease the amount of information the observable manifestations convey. Examples include the administration of potentially invalid questionnaire items in certain contexts (e.g., younger students, lower income countries; Fuchs, 2005; Borgers et al., 2000; Theisen, Achola, & Boakari, 1983), the length of the questionnaire (i.e., few items targeting a specific latent trait; Hox, 2008; Lockwood & McCaffrey, 2014), the number of response categories (Adelson & McCoach, 2010; Simms, et al., 2019; Weng, 2004), and the post hoc manipulation of responses (e.g., rounding, truncation, or categorization; Buonaccorsi, 2010). For instance, questionnaires often include items asking students to identify specific household possessions (e.g., computer, refrigerator, or the material of dwelling floors, walls, and roof). To minimize the burden on students, a questionnaire may limit the number of household possession items or have fewer response categories per item (e.g., two response categories—'yes' or 'no'). Although praiseworthy, efforts to reduce response burden may sacrifice information and introduce undesirable amounts of uncertainty with respect to household possessions or other latent traits (Simms, et al., 2019).

Lastly, creating composites of multiple observable manifestations via factor analysis, item response modeling, or another methodology has favorable measurement qualities (Broer et al., 2019; Caro & Cortes, 2012; Harwell, 2019; May, 2006; Pokropek et al., 2015). Composite measures synthesize information from multiple observable manifestations, are continuous and allow for distributional comparisons (Pokropek et al., 2015), represent a broader range of the latent trait (Cowan et al., 2012), have greater predictive validity than a single manifestation (Lee, Zhang, Stankov, 2019), and are more reliable because they combine information across many observable manifestations (Traynor & Raykov, 2013). Yet, despite their attractiveness, composite measures are error-prone largely because they involve a finite set of imperfect observable manifestations (Hox, 2008; Lockwood & McCaffrey, 2014). As we noted previously, observable manifestations are imprecise and reflect incomplete information; combining them does not eliminate measurement error (Traynor & Raykov, 2013).

Consequences of ignoring measurement error

While there are a variety of conceptualizations and models to examine and address measurement error (e.g., Berkson, multiplicative, differential; Buonaccorsi, 2010; Bound, Brown, & Mathiowetz, 2001), classical additive measurement error is the most common across the social and behavioral sciences. According to Wooldridge (2010) and others (Buonaccorsi, 2010; Carroll et al., 2006; Fuller, 2009; Hausman, 2001), the specification for classical additive measurement error is $\widehat{\theta}_i = \theta_i + u_i$, where $\widehat{\theta}_i$ is the observable

manifestation, θ_i is the true but unobservable latent trait, and u_i is the measurement error for the ith individual. In most cases, we can assume the measurement error has a mean of zero and a constant variance,[4] is conditionally independent of all outcomes and covariates (i.e., nondifferential), and has a nonzero covariance with the observable manifestation and a zero covariance with the true but unobservable latent trait (Buonaccorsi, 2010; Wooldridge, 2010)[5].

The most frequent consequence of ignoring measurement error in an OLS or multilevel regression model is attenuation bias (Gustafson, 2021; Battauz et al., 2011; Wooldridge, 2010; Fuller, 2009; Goldstein, Kounali, & Robinson, 2008; Carroll et al., 2006; Hausman, 2001). For example, in a simple linear regression model where $y_i = \beta_0 + \beta_1\widehat{\theta}_i + \varepsilon_i$ and $\widehat{\theta}_i = \theta_i + u_i$, attenuation bias occurs because the measurement error biases the estimate for $\widehat{\beta}_1$ towards zero. That is, the estimate for $\widehat{\beta}_1$ is the product of the true β_1 and $\sigma^2_{\widehat{\theta}} / \left(\sigma^2_{\widehat{\theta}} + \sigma^2_u\right)$ (which is known as the reliability). Because the reliability for an error-prone manifestation is less than one, $\widehat{\beta}_1$ will always be less than β_1. Lastly, attenuation bias in $\widehat{\beta}_1$ is not the only consequence in this example. Measurement error also biases the estimate for $\widehat{\beta}_0$ and misestimates $\widehat{\varepsilon}_i$ and r^2 (Buonaccorsi, 2010; Carroll et al., 2006; Gustafson, 2021).

The complexity of the consequences intensifies with the inclusion of an error-free covariate, where $y_i = \beta_0 + \beta_1\widehat{\theta}_i + \beta_2 w_i + \varepsilon_i$. If the covariance between the observable manifestation and error-free covariate is zero, the consequences are the same as before; however, if the covariance is nonzero, the measurement error biases the estimate for $\widehat{\beta}_2$ as well (Buonaccorsi, 2010). Further complexities are the inclusion of one or more error-prone manifestations with a nonzero covariance with the observable manifestation (Gustafson, 2021), interactions with error-free covariates or error-prone manifestations (Buonaccorsi, 2010; Yi, 2021), error-prone polynomials (Buonaccorsi, 2010; Yi, 2021), error-prone covariates at multiple levels (Ludtke et al., 2011; Marsh et al., 2009; Televantou et al., 2015), and error-prone outcomes (Buonaccorsi, 2010; Gustafson, 2021; Hausman, 2001). The consequence of the latter is the misestimation of the residual variance in both OLS regression and multilevel regression models (Buonaccorsi, 2010). This is due to the residual variance absorbing the measurement error in the outcome (Gustafson, 2021), which leads to the underestimation of r^2 in OLS regression (Hausman, 2001) and the ICC in multilevel regression models (Battauz et al., 2011; Woodhouse et al., 1996). Outside of error-prone outcomes, the consequences of ignoring measurement error in the previous examples include biasing coefficients towards or away from zero (i.e., attenuation and reverse attenuation), sign reversal, and even order of magnitude reversals among covariates (Gustafson, 2021).[6]

[4] A special case arises when combining observable manifestations via item response theory. The measurement error variance (i.e., σ^2_u) is not constant because the measurement error varies according to the value of the observable manifestation (Battauz & Bellio, 2011; Lockwood & McCaffrey, 2014).

[5] Classical additive measurement error also assumes $\sigma^2_{\widehat{\theta}} = \sigma^2_\theta + \sigma^2_u$ and $\mu_{\widehat{\theta}} = \mu_\theta + 0$. Note that $\sigma^2_{\widehat{\theta}}$ is the variance of the observable manifestation, σ^2_θ is the variance of the true but unobservable latent trait, and σ^2_u is the measurement error variance. The $\mu_{\widehat{\theta}}$ and μ_θ are the means for the observable manifestation and the true but unobservable latent trait.

[6] Riddell's (1989a) study of student achievement among secondary schools in Zimbabwe is a valuable and important contribution to the field of comparative and international education. It is also an example of the consequences of ignoring error-prone covariates (e.g., the grade 7 examination score in models C-D, the quadratic of the grade 7 examination score in model D, and the pupil background index in model D).

Correcting for measurement error

Few studies acknowledge that family background measures are error-prone, that there are consequences to their inclusion in OLS and multilevel regression models, or that there are methods to correct for measurement error. Even if studies acknowledge the error-prone nature of family background measures, they are often unable to correct the measurement error because the software, methods, and auxiliary information are unavailable (Battauz et al., 2011; Buonaccorsi, 2010; Culpepper, 2012). Over the last three decades, software and estimation methods to correct for measurement error have become more accessible to researchers in economics, psychology, sociology, and other fields. These methods include structural equation modeling (e.g., Amos, *lavaan* package in R, Mplus, Stata), instrumental variables (e.g., *ivreg* package in R, Stata), method of moments (e.g., Stata), regression calibration (e.g., Stata), simulation-extrapolation (e.g., Stata, *simex* package in R), and Bayesian estimation (e.g., *brms* and *rstan* packages in R).

The auxiliary information is the most critical because correction for measurement error requires specifying either the reliability, measurement error variance, or the measurement error of the family background measure. For instance, the technical reports from international large-scale assessment programs (e.g., PIRLS, PISA, TIMSS) typically provide an estimate of the reliability for the family background measure (although the reliability will slightly overcorrect regression parameters if it is coefficient α; Culpepper, 2012). The measurement error variance and measurement error are the least accessible auxiliary information, and it is not clear if they are available in the technical reports or databases of international large-scale assessment programs.

Nonetheless, if the reliability, measurement error variance, or the measurement error are known, it is straightforward to demonstrate the correction for measurement error using the previous simple linear model (i.e., $y_i = \beta_0 + \beta_1\widehat{\theta}_i + \varepsilon_i$). Because we know $\widehat{\beta}_1$ equals $\beta_1 \times$ the reliability, then dividing $\widehat{\beta}_1$ by the reliability leaves β_1 (this subsequently corrects $\widehat{\beta}_0$ and accurately estimates $\widehat{\varepsilon}_i$ and r^2; Buonaccorsi, 2010). We can also arrive with β_1 if we know the measurement error variance or the measurement error because, in addition to the variance of the observable manifestation, we will have sufficient information to calculate the reliability. This correction is known as the method of moments correction. In practice, though, correcting an error-prone family background measure is not this simple. Correction will require complex estimation methods given the likely violation of assumptions critical to classical additive measurement error (e.g., measurement error with a nonconstant variance; Lockwood & McCaffrey, 2014).

Methods

The aim of our study is to examine the influence of an error-prone family background measure on the HL Effect in Southern and Eastern Africa. Before discussing our study's methods, we acknowledge that our approach is an intentional simplification of Heyneman and Loxley (1983). The reason for this is our desire to respond to Riddell's claim concerning the reliability of family background measures in lower income countries and the subsequent influence on the HL effect. To do this, we underspecify our models so that they include, within reason, similar student level covariates as those

used by Heyneman and Loxley (1983) to estimate the proportion of variance in student achievement explained by family background (i.e., age in months, gender, and SES). The noticeable difference is our use of an SES composite or index rather than specifying SES as separate, individual covariates as evident in Heyneman and Loxley (1983).[7]

Data sources

We use two data sources in our study. Our primary data source is the SACMEQ III data archive, which includes student achievement and family background data from education systems that participated in the SACMEQ III project between 2005 and 2010. SACMEQ is a regional assessment consortium consisting of 15 ministries of education from Southern and Eastern Africa in partnership with multiple regional universities as well as national and multinational organizations (e.g., UNESCO's International Institute for Educational Planning [IIEP]). The SACMEQ III project is SACMEQ's third assessment cycle (with SACMEQ I in 1995–1998, SACMEQ II in 1998–2004, and SACMEQ IV in 2012–2014).[8] Similar to other international large-scale assessment programs, SACMEQ assessment cycles involve the random selection of schools and students, the administration of tests and contextual questionnaires, and the estimation and scaling of student achievement and questionnaire indices.

To select primary schools and 6th grade students in each education system, the SACMEQ III project employed a two-stage cluster design. The selected students and their respective teachers responded to a questionnaire[9] and took three content area assessments, while leadership at each selected school responded to a questionnaire (Hungi et al., 2010). The entire dataset consists of 61,396 students and 2,779 schools across 15 education systems (i.e., Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Eswatini [formerly known as Swaziland], Tanzania, Uganda, Zambia, Zanzibar, and Zimbabwe). We use three student achievement outcomes (i.e., *zralocp*, *zmalocp*, and *zhalocp*), three student covariates (i.e., *zpagemon*, *psex*, and *pseslog*), the school identifier (i.e., *school*), and the appropriate sampling weight (i.e., *pweight2*) from the SACMEQ III data archive.

The student achievement outcomes are scale scores from three content area assessments administered to students in each education system. The reading, mathematics, and HIV/AIDS knowledge assessments consisted of 55, 49, and 86 items, respectively. The scale scores are maximum likelihood estimates from a Rasch measurement model with a mean of 500 and a standard deviation of 100 (Hungi et al., 2010).[10] The student covariates are *age* (in months), *female* (i.e., pupil gender where female = '1' and male = '0'), and the *SES index* in the logit scale (with a mean of 0 and

[7] Heyneman and Loxley (1983) specified mother's and father's level of education, father's occupation, number of books in the home, and specific household possessions (e.g., dictionary, record player, dishwasher).

[8] More information about SACMEQ is available at http://www.sacmeq.org/.

[9] The SACMEQ III project also provided students with a questionnaire to take home so that parents could verify student responses (Cresswell, Schwantner, & Waters, 2015).

[10] Each SACMEQ III content area assessment was a paper-and-pencil test consisting of a single form with items representing adequate coverage of the construct (Cresswell, Schwantner, & Waters, 2015). All students who participated in SACMEQ III, regardless of education system and school, received the same form and responded to all available items. Consequently, because of the complete item response matrix with approximately zero missing data, the SACMEQ III project could use a Rasch measurement model to calibrate item parameters and estimate student achievement instead of employing a complex methodology (e.g., an item response model with latent regression conditioning on all student assessment and questionnaire responses).

a standard deviation of 1).[11] The *SES index* is a composite of various family background measures[12] via a Rasch measurement model (Dolata, 2008; Hungi et al., 2010). Our rationale for using an index rather than specifying each family background measure as a separate covariate is to take advantage of the availability of the respective measurement error or conditional standard error of measurement (CSEM).[13] As we mentioned previously, the availability of auxiliary information (i.e., reliability, measurement error variance, or CSEM) is vital for correcting an error-prone family background measure. Lastly, our secondary data source is the *World Development Indicators 2009* (World Bank, 2009), which provides us with the 2007 gross national income per capita for the SACMEQ III education systems. The gross national income per capita for 2007 is an estimate using the World Bank's Atlas method, and we refer to it as national income in our analyses for the first and third research questions.

Analysis

Our study consists of several calculations, analyses, and comparisons.[14] First, we calculate descriptive statistics using the *survey* package in R (Lumley, 2021) and the appropriate sampling weight. The descriptive statistics for each SACMEQ III education system are found in Tables 1 and 2.[15] These include national income in 2007, count of students and schools in the sample, and the estimates of the mean and respective standard error for each outcome (i.e., reading, mathematics, HIV/AIDS knowledge achievement) and covariate (i.e., *age*, *female*, and the *SES index*). We also estimate the variance for the *SES index* and CSEM as well as the reliability for the *SES index*. To address the first research question, we plot the reliability for the *SES index* and national income for all SACMEQ III education systems.

Second, we estimate six multilevel regression models for each SACMEQ III education system using Bayesian estimation via the *brms* package in R (version 2.16.3; Burkner, 2017) and *rstan* package in R (version 2.21.3; Stan Development Team, 2021).[16] Our models rely on minimally informative priors similar to Lockwood and McCaffrey (2014) and Junker, Schofield, and Taylor (2012).[17] The six models consist of a naive and

[11] The unit of measurement for the Rasch measurement model is the logit scale or the logarithm of the odds (de Ayala, 2009). The *SES index* in a logit scale (i.e., *pseslog*) is not the official SACMEQ III reporting scale. The SACMEQ III project applied a transformation to *pseslog* to produce *zpsesscr*. We use *pseslog* rather than *zpsesscr* because the CSEM shares the same scale as *pseslog* and *rstan* runs faster with standardized covariates.

[12] The *SES index* uses student responses to 18 items relating to parent education, number of books in the home, specific household possessions, source of dwelling lighting, and the material of dwelling floors, walls, and roof.

[13] The name of the CSEM for the *SES index* in the SACMEQ III data archive is *psesse*. Because the *SES index* is a maximum likelihood estimate, the CSEM for the jth *SES index* is the inverse square root of the test information function (TIF) (Baker & Kim, 2017). Moreover, the TIF is the sum of $I_i\left(\widehat{\theta}_j\right) = a_i^2 P_i Q_i$ (Baker & Kim, 2017). Note that $\widehat{\theta}_j$ represents the *SES index* for the jth student, $I_i\left(\widehat{\theta}_j\right)$ is the information function for the ith item and respective *SES index*, a_i^2 is the square of the discrimination parameter for the ith item, P_i is the probability of a response with the highest value for the ith item, and Q_i is $1 - P_i$.

[14] All analyses rely on R version 4.1.2 (R Development Core Team, 2021).

[15] We do not report the count or percent of cases with missing data in Tables 1 and 2 because missingness impacts less than one percent of cases. All cases have complete data for the covariates and reading achievement, while 99.9 and 99.8 percent of cases have complete data for mathematics and HIV/AIDS knowledge achievement, respectively.

[16] We use Bayesian estimation because the measurement error variance is heterogeneous given the observable *SES index* is an estimate from an item response model. With respect to the incorporation of sampling weights within Bayesian estimation, the *brms* package multiplies each case's log-posterior by their respective sampling weight.

[17] We use minimally informative priors because we are uncertain about the association between SES and student achievement in lower income countries after correcting for measurement error. There are several meta-analyses exploring a naive association in wealthy countries (Harwell et al., 2017; Sirin, 2005; White, 1982) and in lower income countries (Kim, Cho, & Kim, 2019). The association seems rather small in Sub-Saharan African countries per Kim, Cho, and Kim (2019); however, we are unsure of the magnitude and direction of the association after applying a correction for

Table 1 Education system characteristics and achievement descriptive statistics

Education system	National income	Sample		Reading achievement		Mathematics achievement		HIV/AIDS knowledge achievement	
		Students	Schools	M	SE	M	SE	M	SE
Botswana	6,120	3868	160	534.64	4.96	520.53	3.78	499.10	4.65
Eswatini	2,560	4030	172	549.39	2.98	540.84	2.37	530.80	3.19
Kenya	640	4436	193	543.11	5.21	556.96	4.19	509.27	4.64
Lesotho	1,030	4240	182	467.87	3.03	476.91	2.74	464.74	3.79
Malawi	250	2781	139	433.50	2.60	447.02	2.84	511.70	6.47
Mauritius	5,580	3524	152	573.54	5.46	623.27	6.39	453.09	4.67
Mozambique	330	3360	183	476.04	3.56	483.81	2.67	507.00	6.15
Namibia	3,450	6398	267	496.92	3.91	471.03	2.99	502.04	3.81
Seychelles	8,960	1480	24	575.10	6.55	550.65	5.14	487.72	4.82
South Africa	5,720	9071	392	494.95	5.41	494.84	4.27	502.82	4.18
Tanzania	410	4194	196	577.76	3.49	552.72	3.61	575.90	4.45
Uganda	370	5307	264	478.68	3.65	481.90	3.11	489.01	4.33
Zambia	770	2895	157	434.41	3.44	435.15	2.51	488.09	4.46
Zanzibar	410	2791	143	533.88	3.54	486.20	2.43	500.84	3.37
Zimbabwe	340	3021	155	507.68	6.68	519.79	5.68	477.27	5.72

Note. National income is the 2007 gross national income per capita using the World Bank Atlas method (World Bank, 2009). M represents the sample mean and SE is the respective standard error. All calculations of sample means and standard errors use the *survey* package in R, version 4.1-1 (Lumley, 2021)

Table 2 Student and family background descriptive statistics

Education system	Age (in Months)		Females		SES Index		Variance		Reliability ratio
	M	SE	M	SE	M	SE	SES Index	CSEM	
Botswana	153.45	0.40	0.50	0.01	0.47	0.07	1.47	0.29	0.84
Eswatini	166.25	0.57	0.50	0.01	0.16	0.05	1.05	0.23	0.82
Kenya	165.14	0.80	0.49	0.01	−0.62	0.05	0.79	0.20	0.80
Lesotho	168.02	0.64	0.55	0.01	−0.50	0.04	0.69	0.19	0.78
Malawi	169.47	0.86	0.49	0.01	−1.01	0.04	0.74	0.23	0.76
Mauritius	136.45	0.14	0.49	0.01	1.55	0.04	0.96	0.45	0.68
Mozambique	170.21	0.96	0.46	0.01	−0.90	0.06	1.17	0.27	0.81
Namibia	163.21	0.57	0.52	0.01	−0.29	0.06	1.31	0.23	0.85
Seychelles	138.59	0.09	0.49	0.01	1.46	0.04	0.52	0.37	0.58
South Africa	154.56	0.39	0.51	0.01	0.65	0.05	1.37	0.30	0.82
Tanzania	174.98	0.81	0.51	0.01	−1.03	0.04	0.62	0.22	0.74
Uganda	169.41	0.68	0.51	0.01	−0.91	0.04	0.74	0.22	0.77
Zambia	168.99	0.93	0.49	0.01	−0.65	0.06	1.08	0.23	0.82
Zanzibar	169.70	0.69	0.57	0.01	−0.53	0.06	0.88	0.21	0.81
Zimbabwe	149.20	0.48	0.56	0.01	−0.22	0.07	1.03	0.21	0.83

Note. M represents the sample mean and SE is the respective standard error. All calculations of sample means, standard errors, and variances use the *survey* package in R, version 4.1-1 (Lumley, 2021)

Footnote 17 (continued)

measurement error. Thus, we use a *normal(0, 100)* prior for the intercepts, slopes, and the latent mean and standard deviation for the true *SES index* (i.e., μ_θ and σ_θ), and a *cauchy(0, 100)* prior for the random effects.

correction version for each content area (i.e., reading, mathematics, and HIV/AIDS knowledge). For each parameter in all models, we use four Markov chains, 1,000 warm-up iterations, and 6,000 total iterations. This results in a posterior distribution with four × 5,000 or 20,000 samples per parameter. After estimating the posterior distribution, we evaluate convergence by using the Gelman-Rubin statistic (Gelman & Rubin, 1992) and visually inspecting the trace and density plots for each parameter as well as graphical posterior predictive checks.[18]

The naive model is a random intercept model where $y_{ij} = \mu_{00} + \beta_{1j}Age_{ij} + \beta_{2j}Female_{ij} + \beta_{3j}\widehat{\theta}_{ij} + \sigma_{ij} + \sigma_{0j}$, and assumes all covariates are error-free (even the *SES index* or $\widehat{\theta}_{ij}$). The product of the random intercept model and the priors is a posterior distribution comprising of 20,000 samples for each parameter (i.e., μ_{00}, β_{1j}, β_{2j}, β_{3j}, σ_{ij}, and σ_{0j}) and each school's intercept. On the other hand, the correction model acknowledges the error-prone nature of the *SES index* and corrects it by including a measurement model to estimate the latent but true *SES index* (or θ_{ij}). The measurement model is $\widehat{\theta}_{ij} \sim Normal\left(\theta_{ij}, u_{ij}\right)$, and we model the observable *SES index* has having a normal distribution with an unknown center (i.e., the true *SES index*) and a known standard deviation (i.e., the CSEM or u_{ij}). The random intercept model is now $y_{ij} = \mu_{00} + \beta_{1j}Age_{ij} + \beta_{2j}Female_{ij} + \beta_{3j}\theta_{ij} + \sigma_{ij} + \sigma_{0j}$, where θ_{ij} is the true *SES index* for the ith student in the jth school. Thus, with respect to the correction model, our aim is to estimate a posterior distribution with 20,000 samples for the true *SES index*, the other model parameters (i.e., μ_{00}, β_{1j}, β_{2j}, β_{3j}, σ_{ij}, and σ_{0j}), the latent mean and standard deviation for the true *SES index* (i.e., μ_θ and σ_θ), and the intercept for the jth school.

To address the second and third research questions, we plot multiple associations: (1) the naive and correction parameter estimates for the *SES index* by content area and (2) the school effect (or σ_{0j}) and national income for SACMEQ III education systems by naive and correction models. Although we use six figures to explore the sensitivity of the HL effect to measurement error, Additional file 1A contains the naive and correction parameter estimates by SACMEQ III education system and content area assessment (i.e., Tables S1 through S9). Moreover, trace and density plots, graphical posterior predictive checks, model output by SACMEQ III education system, and R code are available upon request.

Assumptions and limitations

Our assumptions prior to conducting the analysis are sixfold. First, we assume *Age* and *Female* are error-free. We believe this is a tenable assumption although we are aware of a few studies suggesting *Age* may be error-prone (Bedard & Dhuey, 2006; Lee & Fish, 2010). Second, we assume *Age*, *Female*, and the *SES index* have covariances equal to or near zero. We have no evidence of meaningful collinearity, but we acknowledge the absence of collinearity is improbable. *Age*, *Female*, and the *SES index* have nonzero covariances; however, we believe these covariances are negligible. Third, similar to Chudgar and Luschei (2009) and Riddell (1989a), we do not specify covariates representing school quality. This intentional specification requires us to assume the school effect only reflects school quality rather than other school, community, or student characteristics.

[18] For each naive and correction model, the Gelman-Rubin statistic (i.e., $\widehat{R}$) has a value of 1 suggesting adequate convergence, and the trace plots by parameter show random scatter or mixing across iterations and chains.

We acknowledge this assumption may not be defensible given the influence of school composition effects (Marsh et al., 2012; Thrupp et al., 2002).

Fourth, we assume the three outcomes are error-free. This assumption is clearly erroneous because they are estimates from a Rasch measurement model and represent observable but error-prone reading, mathematics, and HIV/AIDS knowledge achievement. Nonetheless, as we mentioned earlier, ignoring error-prone outcomes does not bias parameter estimates but does underestimate the ICC in multilevel regression models (Battauz et al., 2011; Woodhouse et al., 1996). The error-prone nature of the outcomes is unlikely to impact our analysis and findings; however, we believe this is a promising area for future research especially in the context of the HL Effect.[19] Fifth, we recognize the intentional underspecification of our models is a limitation. The inclusion of relevant student and school level covariates (e.g., additional family background measures, educator, and school quality measures) may alter our findings, shed light on findings we did not anticipate, or even contradict theoretical claims and prior research. We intend to expand our specification as part of future research with covariates similar to recent studies exploring the HL Effect (e.g., Bouhlila, 2015; Chiu, 2007; Chudgar & Luschei, 2009).

Sixth and finally, we concede our findings may not be generalizable beyond the education systems that are part of the SACMEQ III project. While the student and school samples are reasonably large (i.e., 61,396 students and 2779 schools), the sample of 15 education systems is small and homogenous in comparison to the number and variety of countries participating in other international large-scale assessment programs. Nonetheless, the advantage of using the SACMEQ III data archive over PIRLS, PISA, or TIMSS is the availability of the CSEM for the *SES index*. Since the aim of our study is to examine the HL Effect and the influence of an error-prone family background measure, we believe trading external validity for access to the CSEM is worthwhile. Yet, we do not wish to minimize this limitation. Exploring the sensitivity of the HL Effect to measurement error across a large number of culturally, economically, and geographically diverse countries would be a substantial contribution. Thus, we strongly encourage a replication of our current study using other international large-scale assessment programs.[20]

[19] The assessments used as outcomes by Heyneman and Loxley (1983) may have been more error-prone than the family background measures. For instance, the science assessment used in Argentina, Bolivia, Brazil, Colombia, Mexico, Paraguay, and Peru relied on a subset of science items from the First International Science Study (FISS). This subset represented approximately half the test length of the FISS assessment and consisted of the easiest items (see footnote 16 in Heyneman and Loxley [1983] for more details). We know that test length and item information determine reliability; thus, this suggests that seven of the 16 countries with the lowest national income in Heyneman and Loxley (1983) had an outcome with a lower reliability than the wealthiest countries.

[20] An analysis using the PISA 2009 database would represent a promising replication study. The design of the replication would include (1) the estimation of the CSEM for the household possessions index via recalibrating the item parameters and re-estimating the household possessions index for each student, and (2) the estimation of a naive and correction Bayesian multilevel model for each plausible value, content area (i.e., math, reading, and science), and country. For planning purposes, this would result in $2 \times 5 \times 3 \times 75$ or 2250 posterior distributions, where the estimation of each posterior distribution would require 5 to 10 min of computation time for a computer with 16 gigabytes of memory. Computation time is certainly not insurmountable but we recognize this could be a heavy lift for those who wish to replicate our study using PISA 2009 or data from other international large-scale assessment programs.

Results and discussion

Our study explores the sensitivity of the HL Effect to an error-prone family background measure in Southern and Eastern Africa via three research questions. The first question focuses on the association between national income and the reliability for the *SES index* in SACMEQ III education systems. *Does national income have an association with the reliability of a family background measure?* This association is visible in Fig. 1, and appears to be negative ($r = -0.49$). Interestingly, if we remove the two education systems with the lowest reliability for the *SES index* (i.e., Mauritius and Seychelles), the association becomes positive ($r = 0.57$).[21] The change in direction is entirely due to the small number of education systems, where the omission of one or two extreme values alters the direction. While we observe a negative association between national income and the reliability for the *SES index* in SACMEQ III education systems, we acknowledge the challenges with examining associations with small n-sizes and influential outliers; thus, we urge caution and restraint with interpretations and inferences. We wonder, however, if this association would remain, strengthen, or even reverse if our sample of education systems were larger and included more regional neighbors like Angola, Burundi, Ethiopia, Madagascar, Rwanda, or others (e.g., the 10 West African countries participating in the Program for the Analysis of Education Systems [PASEC]). We are also curious what the observable association is outside of SACMEQ III education systems and Sub-Saharan Africa.[22] Thus, we encourage and invite further research into the association between national income and the reliability of family background measures in culturally, economically, and geographically diverse countries.

For the sake of discussion, suppose the negative association between national income and the reliability of the *SES index* is true and generalizable. First, we must admit the observable pattern in the SACMEQ III data did not match our a priori expectations or our interpretation of Riddell's claim. We expected family background measures to have a higher reliability in wealthier countries than in lower income countries. To some extent, this finding is baffling. Has this always been the case or is this a new phenomenon? Why would the reliability be lower in wealthier countries? One way to approach this is to consider the calculation of the reliability for the *SES index*, $\sigma^2_{\hat{\theta}} / \left(\sigma^2_{\hat{\theta}} + \sigma^2_u \right)$. If the measurement error variance remains constant, the reliability will increase as long as the variance of the *SES index* increases. That is, the reliability increases when the variance in the *SES index* increases (and if the differences due to measurement error variance are constant). Let us assume measurement error variance is constant and equal across wealthy and lower income countries. A lower reliability in wealthy countries would suggest smaller differences between students in terms of the *SES index*. Are students

[21] The p-values for r = -.49 and r = .57 are p = .063 and p = .044, respectively.

[22] We provide the results from a preliminary study of PISA 2009 in Fig. S1 of Additional file 1B. We observe a negative association between national income and the reliability of the household possessions index in PISA 2009 countries. This potentially corroborates what we find with SACMEQ III education systems; however, it is important to note that PISA 2009 uses coefficient α to estimate the reliability of the household possessions index (OECD, 2012). As noted by Culpepper (2012) and Sijtsma (2009), coefficient α is an estimate of the lower bound of reliability (and possibly a serious underestimate of reliability). The distinct reliability measures in Figs. 1 and S1 and their respective associations with national income are weakly comparable; thus, we urge caution with interpretations and inferences.

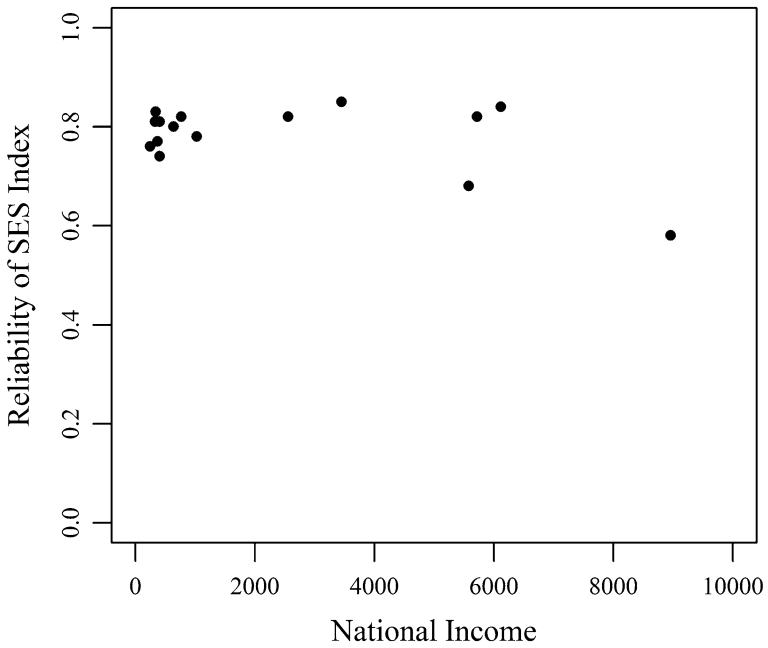

Fig. 1 National income and the reliability of the SES Index in SACMEQ III education systems

more homogeneous in wealthy countries? Perhaps. A more probable explanation is the items which comprise the *SES index* represent manifestations of family background that have become ubiquitous in wealthier countries. Questionnaire items with homogeneous and extreme response patterns (e.g., 'yes' to all household possessions) contribute little to reliability (Traynor & Raykov, 2013).[23]

Returning to Riddell's claim and the HL Effect, a negative association between national income and the reliability of a family background measure would have profound implications. Under our expectation and Riddell's claim, correcting for measurement error in Heyneman and Loxley (1983) would disattenuate the influence of family background and diminish the percent of variance in student achievement attributable to school quality more in lower income countries than in wealthy countries. The consequence would be a weaker or negligible association between national income and school quality. On the other hand, if the association were reverse as we observe in SACMEQ III education systems, a correction for measurement error would likely strengthen the HL Effect because the percent of variance in student achievement attributable to school quality would shrink more in wealthy countries than in lower income countries.

[23] Another explanation for a lower reliability in wealthy countries is missing item responses for the family background measure. In general, the measurement error for a family background measure will be larger in the presence of missing data because incomplete item response patterns provide less information. Thus, the reliability of a family background measure will be lower in wealthy countries if students from those respective countries are more likely to have missing item responses than students from lower income countries. In the context of the SACMEQ III project, all education systems except Zimbabwe had 99 percent or more students with complete item response patterns for the *SES index*. While missing item responses did not influence the reliability of the *SES index*, missing item responses may influence the reliability of family background measures in wealthy or lower income countries participating in other international large-scale assessment programs.

The second research question focuses on the measurement error sensitivity of the association between the *SES index* and student achievement in SACMEQ III education systems. *Is the association between a family background measure and student achievement sensitive to measurement error?* Figures 2, 3, and 4 display the naive and correction *SES index* parameter estimates for reading, mathematics, and HIV/AIDS knowledge achievement. The x-axes are the *SES index* parameter estimates from the naive models and the y-axes are the *SES index* parameter estimates from the correction models. Both the naive and correction parameter estimates are the posterior means.[24] Points directly on the identity lines imply no sensitivity to measurement error (i.e., the correction and naive parameter estimates are identical), while points falling above or below the identity lines represent sensitivity to measurement error. Sensitivity is the difference between the naive and correction parameter estimates, either as an increase/decrease in raw scale score points or as a percentage change. According to Figs. 2, 3, and 4, the association is sensitive to measurement error as all points by content area assessment and SACMEQ III education system are above the identity line (reflecting attenuation bias since the naive *SES index* parameter estimates are smaller than those from the correction model).

The sensitivity ranges from approximately 5 to 161 scale score points for reading, 2–138 scale score points for mathematics, and 4–90 scale score points for HIV/AIDS knowledge. Across content areas, Mauritius, Seychelles, Tanzania, and Zimbabwe are the most sensitive to measurement error with the largest increases in raw scale score points, while Eswatini, Malawi, and Uganda are the least sensitive. Most SACMEQ III education systems experience an increase from the naive to the correction parameter estimates of greater than 100 percent in all content areas (except for Botswana, Namibia, and Eswatini). For example, Botswana has the smallest percentage increases with 49, 54, and 57 percent in reading, mathematics, and HIV/AIDS knowledge. The opposite of Botswana is Seychelles with 323, 334, and 345 percent increases in the respective content areas. An interesting case is Uganda with a 610 percent increase in the association between the *SES index* and mathematics achievement. The reason Uganda is interesting is because the raw scale score increase between the naive and correction parameter estimates is only 4.76 points. Whether by raw scale score points or percentage changes, it is clear the association between the *SES index* and student achievement is sensitive to measurement error. Of course, the sensitivity is not uniform across SACMEQ III since some education systems experience more sensitivity than others—this certainly depends on the respective reliability and the association between the *SES index* and student achievement.

At first glance, the sensitivity across education systems appears quite sizable. Yet, it is important to remember that the logit scale of the *SES index* is logarithmic. Rather than the typical interpretation (i.e., a one unit increase in x corresponds to a β unit change in y), the interpretation of a slope for a covariate with a logarithmic scale is a $\beta \times \ln(1.01)$ unit change in y corresponding to a one percent change in x. Using the correction parameter estimates for reading in Uganda as an example, a 25 percent change in the *SES index* results in a $17.24 \times ln(1.25)$ or 3.85 change in reading scale score points (after adjusting for the other covariates and random effects). Though a 25 percent change in the *SES index* is considerable, a 3.85 change in reading scale score points is quite small

[24] The posterior means and standard deviations are available in Tables S1 through S9 in Additional file 1A.

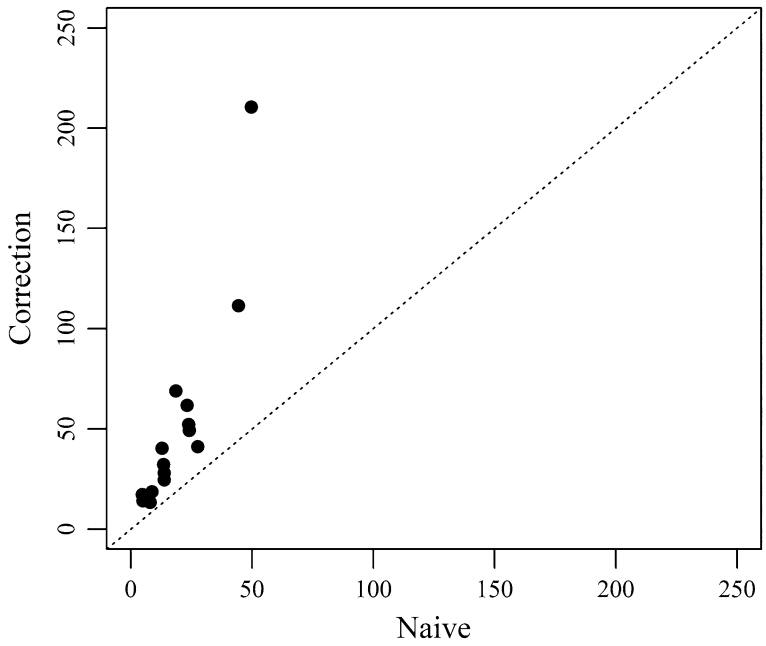

Fig. 2 Reading naive and correction model parameters for the SES Index. Note. Parameter estimates are the posterior mean

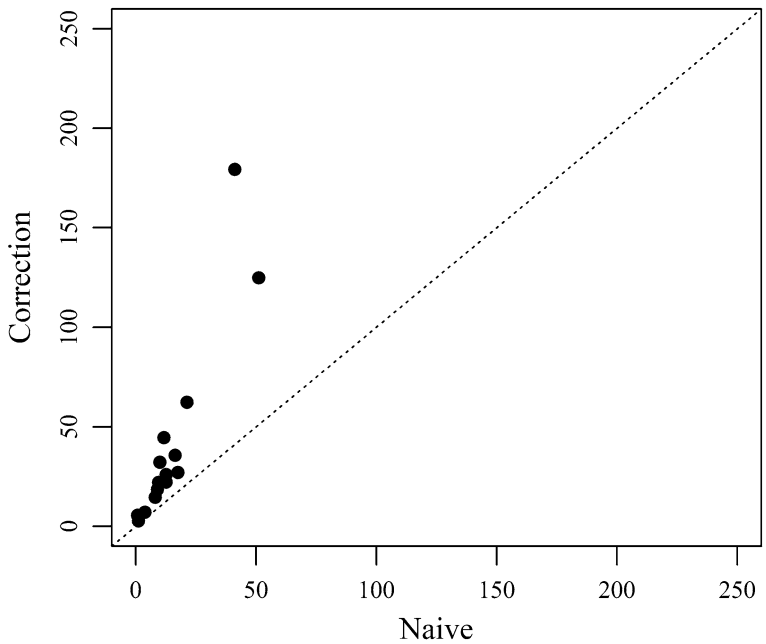

Fig. 3 Mathematics naive and correction model parameters for the SES Index. Note. Parameter estimates are the posterior mean

given a standard deviation of 100 for the reading scale. If we interpret each SACMEQ III education system's correction parameter estimates as influencing a 25 percent change in reading, mathematics, or HIV/AIDS knowledge achievement, all produce minor changes in scale score points (ranging from 0.62 in Malawi for mathematics to 15.38 in

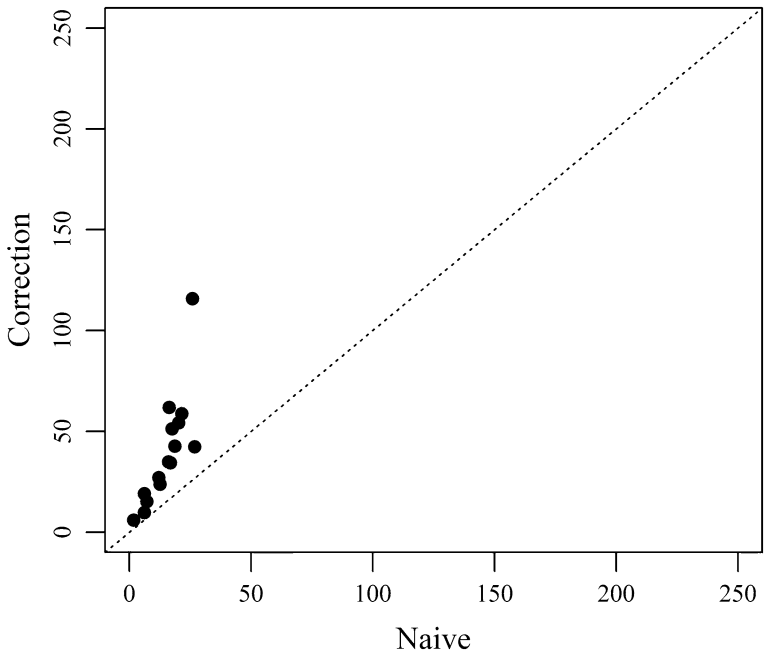

Fig. 4 HIV/AIDS knowledge naive and correction model parameters for the SES index. Note. Parameter estimates are the posterior mean

Tanzania for reading) except for Mauritius and Seychelles (e.g., 27.83 in Mauritius for mathematics and 46.97 in Seychelles for reading). While the correction for measurement error in the *SES index* yields larger values for the true *SES index* parameters, the larger values still reflect a rather weak association with reading, mathematics, and HIV/AIDS knowledge achievement. Mauritius and Seychelles are exceptions. The reason for this are their low reliability and higher naive parameter estimates for the *SES index*.

Although the associations between the *SES index* and student achievement are sensitive to measurement error (and increase after correction), the magnitude of the associations remain small in nearly all of the SACMEQ III education systems. This has interesting implications for the HL Effect. We believe the family background measures in the original Heyneman and Loxley study were error-prone and their associations with science achievement were sensitive to measurement error. What is unknown is the magnitude of the sensitivity. Looking at Table 2 on pages 1174–1175 of Heyneman and Loxley (1983) and among countries with national incomes less than $1,000, we believe only the association for Paraguay would meaningfully change with correction (assuming a low reliability of 0.65). The reason for Paraguay, and no other countries with similar or lower national incomes, is its second largest value for the percentage of variance explained by preschool influences among all countries in Heyneman and Loxley (1983). The correction for measurement error in countries like Colombia, India, or Uganda would result in a meaningless change given their small values for the percentage of variance explained by preschool influences (i.e., 1.8, 2.7, and 5.8). The lesson here is no amount of measurement error correction can make up for an already small, weak, or negligible association with student achievement (unless the reliability is absurdly low;

e.g., less than 0.10). Thus, we wonder if correcting for error-prone family background measures in Heyneman and Loxley (1983) would significantly alter the conclusions.

The third and final research question focuses on the association between national income and the school effect, and whether that association is sensitive to measurement error in Southern and Eastern Africa. *Is the association between national income and the school effect sensitive to measurement error?* This association is visible in Figs. 5, 6, and 7 for reading, mathematics, and HIV/AIDS knowledge achievement. The x-axes are national income, the y-axes are the school effect, and the black points and gray triangles represent the naive and correction parameter estimates, respectively. The school effect is the standard deviation of school intercepts, where larger values imply differences between schools in terms of average achievement (or, in other words, school quality; Chudgar & Luschei, 2009). The HL Effect is the negative association between national income and the school effect. As seen in Figs. 5, 6, and 7, the associations between national income and the school effect are reasonably insensitive to measurement error across content areas assessments. That is, there is substantial overlap between the naive and correction parameter estimates (i.e., points and triangles) resulting in approximately identical associations between national income and the school effect.

Although the associations are insensitive to measurement error, the school effect in some SACMEQ III education systems are more sensitive than others. For instance, the sensitivity in terms of raw scale score points ranges from 0.1 to 5.1 in reading, approximately zero to 4.4 in mathematics, and approximately zero to 2.3 scale score points in HIV/AIDS knowledge achievement. In each content area, Zimbabwe is the most sensitive, while Eswatini is the least sensitive in reading and mathematics, and Mauritius is the least sensitive in HIV/AIDS knowledge. The percentage increase between the naive and correction school effect yields comparable results; Zimbabwe is the most sensitive with percentage changes of 9.1, 9.4, and 4.5 percent in reading, mathematics, and HIV/AIDS knowledge, respectively. Albeit quite small, we did not anticipate observing differences between the naive and correction school effect in any of the SACMEQ III education systems. These differences may suggest the error-prone *SES index* is level-2 endogenous (Bates et al., 2014; Castellano et al., 2014); meaning it has a nonzero covariance with the school effect and the CSEM in the *SES index* slightly biases the naive school effect.

Lastly, the associations between national income and the school effect for reading and mathematics achievement echo the findings from contemporary studies; the HL effect is vanishing (Baker et al., 2002; Bouhlila, 2015; Chmielewski, 2019). That is, for reading and mathematics achievement, we do not observe a negative association between national income and the school effect (which reflects the Coleman Effect rather than the HL Effect). On the other hand, the association we observe for HIV/AIDS knowledge is clearly negative, implying the quality of schooling has a greater influence on HIV/AIDS knowledge achievement in lower income SACMEQ III education systems ($r_{Naive} = -0.56$; $r_{Correction} = -0.55$).[25] Similar to the findings from our first research question, we urge

[25] The p-values for r = -.56 and r = -.55 are p = .029 and p = .032, respectively

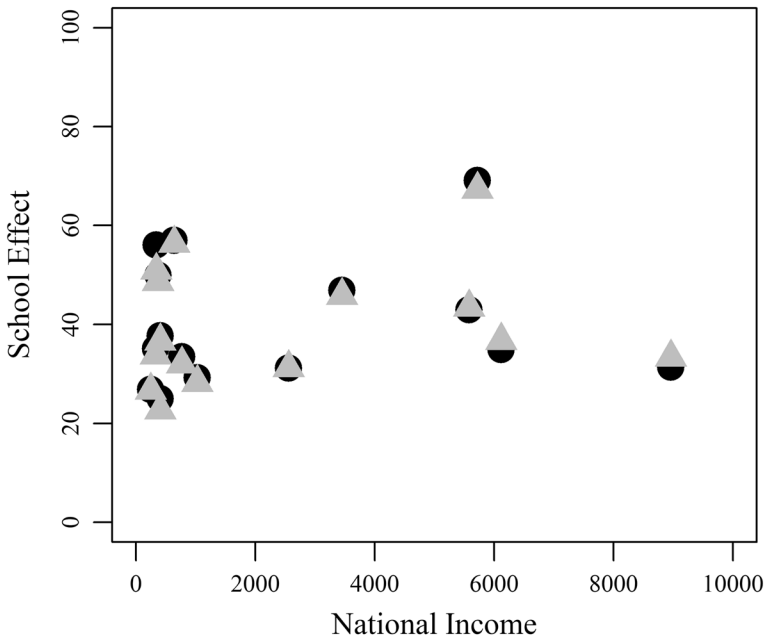

Fig. 5 National income and the school effect for reading. Note. Black points and gray triangles are the naive and correction parameter estimates for the school effect. Both school effects are posterior means

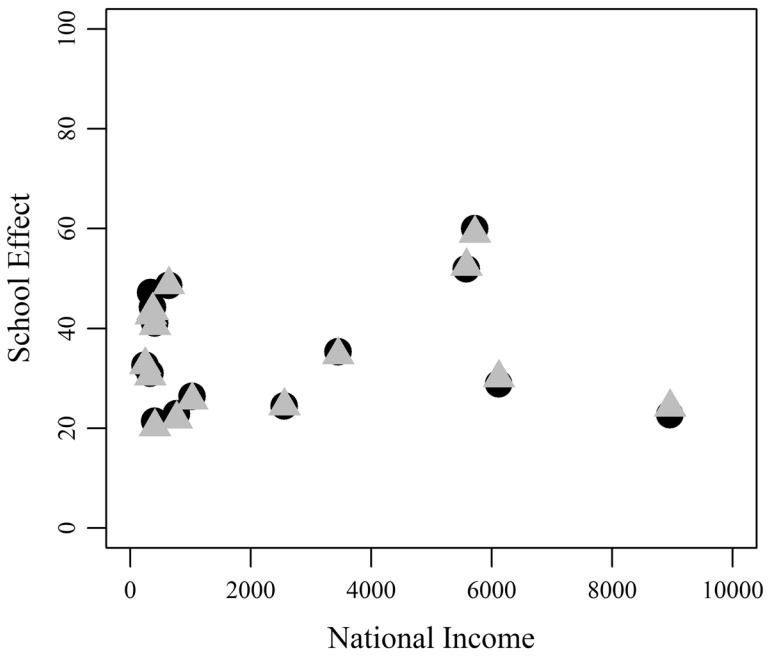

Fig. 6 National income and the school effect for Mathematics. Note. Black points and gray triangles are the naive and correction parameter estimates for the school effect. Both school effects are posterior means

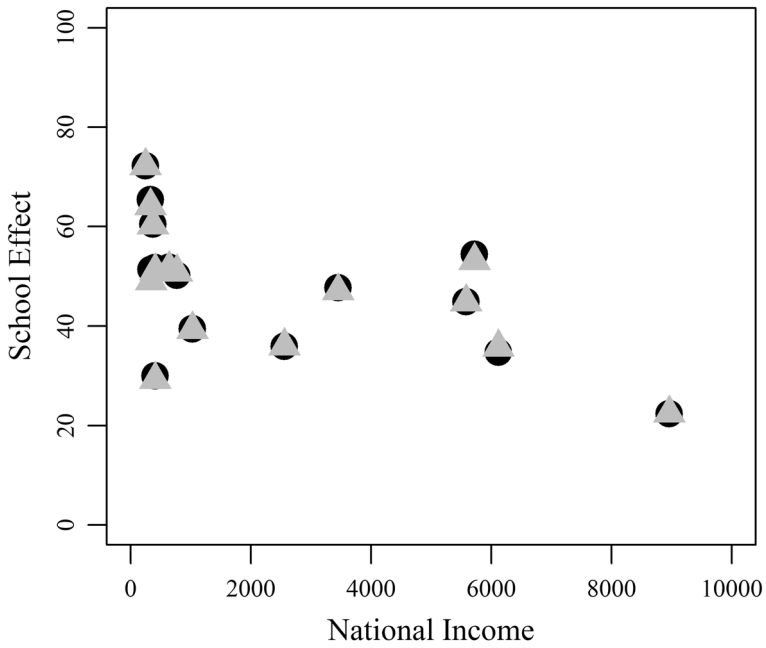

Fig. 7 National income and the school effect for HIV/AIDS Knowledge. Note. Black points and gray triangles are the naive and correction parameter estimates for the school effect. Both school effects are posterior means

caution with interpretations and inferences given the small n-size.[26] Nonetheless, we did not foresee this association. Much of the research exploring the HL Effect relies on reading, mathematics, and science achievement. This is due to the availability of international large-scale assessment programs (e.g., PISA, PIRLS, and TIMSS), which emphasize reading, mathematics, and science achievement over other content areas. The SACMEQ III project and its inclusion of HIV/AIDS knowledge is genuinely unique in comparison to other international large-scale assessment programs.

Why is the HL Effect present in HIV/AIDS knowledge but not in reading or mathematics across the SACMEQ III education systems? We can only speculate but a plausible explanation may be differential access to information concerning HIV/AIDS in lower income versus wealthier SACMEQ III education systems. Families, regardless of background, may be more reliant on schools for information about HIV/AIDS in lower income education systems (e.g., Malawi, Mozambique, and Uganda) because basic public health institutions are weak or not available. This means the quality of schools may matter more than family background in the absence of public health institutions. In contrast, this also implies, similar to the assertion by Baker et al. (2002), the mass expansion of basic public health institutions will inevitably erode the influence of school quality relative to family background (with respect to HIV/AIDS knowledge achievement). An example of this may be Seychelles, which has a free and universal basic healthcare system. From our analyses, Seychelles has the largest and strongest

[26] We examined the influence of outliers by estimating the correlations after removing SACMEQ III education systems with the largest values for the school effect (i.e., Malawi and Mozambique) and national income (i.e., Botswana and Seychelles). The correlations remained negative without Malawi and Mozambique ($r_{Naive} = -.50$; $r_{Correction} = -.49$) or remained negative but weakened without Botswana and Seychelles ($r_{Naive} = -.21$; $r_{Correction} = -.21$).

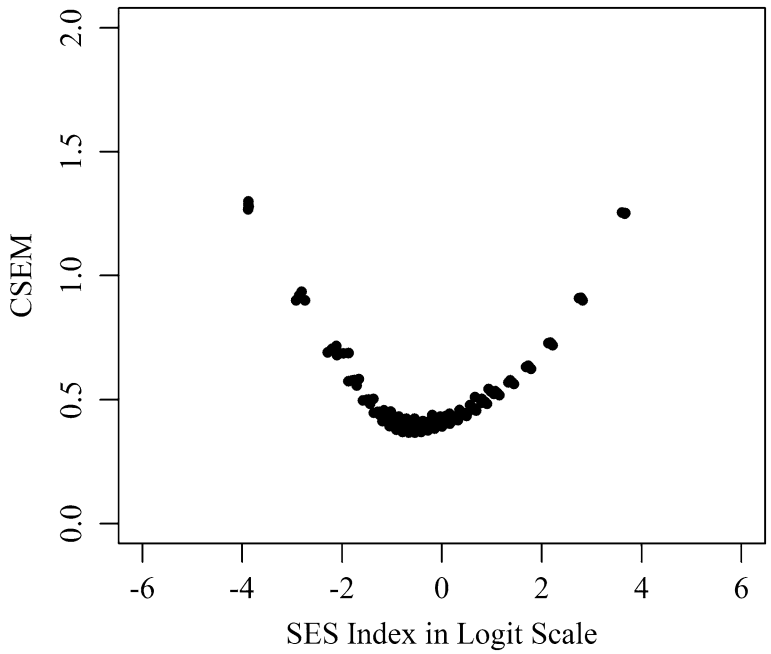

Fig. 8 SES Index and the CSEM for All SACMEQ III education systems

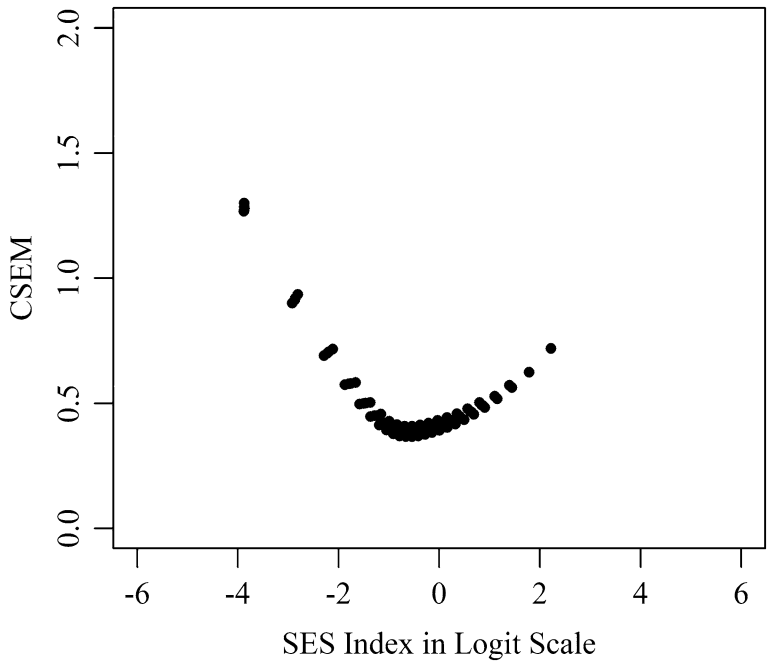

Fig. 9 SES Index and the CSEM for Malawi

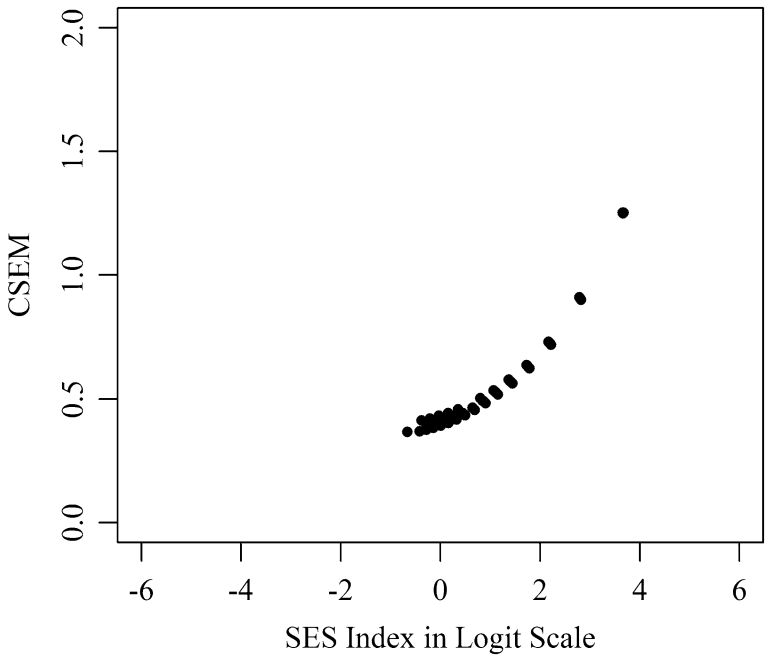

Fig. 10 SES Index and the CSEM for Seychelles

association between the *SES index* and HIV/AIDS knowledge achievement, and the smallest school effect for HIV/AIDS knowledge; all suggesting a vanishing HL Effect.

Conclusion

Our study investigates the sensitivity of the HL Effect to an error-prone family background measure in 15 SACMEQ III education systems. We revisit Riddell's claim concerning the reliability of family background measures and its respective association with national income, and explore the sensitivity of two associations to measurement error (i.e., family background with student achievement and national income with the school effect). Some of our findings contradict our expectations and the claims from prior research, while others provide an interesting and unique perspective on the HL Effect (notwithstanding its presence or absence). Regardless of our findings, we wish to underscore the importance of correcting for measurement error or, at the very least, acknowledging its existence and influence when correction is unavailable. This is particularly relevant to revisiting or challenging extant theories and effects (e.g., Coleman and HL Effects). We concur with Loken and Gelman (2017) concerning measurement error and scientific replication, and believe any study attempting to replicate the HL Effect must correct error-prone covariates (and, as we mentioned earlier, outcomes). Lastly, our findings do not suggest families or schools are irrelevant in Southern and Eastern Africa (or in any country for that matter). If anything, our findings convey the difficulty and complexity in determining the role of families and schools, their influence on student achievement, and the uniformity of this influence across countries. Recognizing these complexities raises our appreciation of Coleman et al. (1966), Heyneman and Loxley (1983), and Riddell (1989a), their respective methodological and theoretical contributions, and being an inspiration for subsequent studies.

Throughout this study, we exclusively focus on the correction of error-prone family background measures and the HL Effect. An unintentional corollary of our correction of the SACMEQ III *SES index* is the association between the *SES index* and the CSEM. Besides confirming that the measurement error variance is heterogeneous given the CSEM varies according to each value of the *SES index*, plotting the association illustrates where the uncertainty is across the *SES index* distribution. Figures 8, 9, and 10 display the association for all SACMEQ III education systems, Malawi, and Seychelles. The considerable uncertainty at the ends of the *SES index* distribution is evident for all education systems in Fig. 8, which suggests weak item targeting for low and high SES students. That is to say, the *SES index* lacks items that either represent manifestations of low or high SES or are informative at the ends of the SES distribution. The patterns are reasonably distinct and interesting for Malawi and Seychelles. We observe more uncertainty and weak item targeting for low SES in Malawi (Fig. 9) and the reverse in Seychelles (Fig. 10). We do not believe this is a coincidence. The *SES index* and CSEM associations behave as reasonable opposites in Malawi and Seychelles because of the stark difference in national income ($250 versus $8,960). What we observe in Malawi and Seychelles seems unique, but weak item targeting may be the rule across countries than the exception.

If true, this has profound methodological and practical implications concerning the development, interpretation, and use of family background measures. Improving the reliability of the *SES index*, and other family background measures, will require better item targeting and, more than likely, the inclusion of items reflecting local context. Of course, concerns about the reliability and comparability of family background measures are not new (Rutkowski & Rutkowski, 2013, 2018; Sandoval-Hernandez et al., 2019). To some extent, it feels like a tradeoff; improving reliability comes at the expense of comparability and vice versa. As noted by Buchmann (2002), "the challenge is to walk the fine line between sensitivity to local context and the concern for comparability across multiple contexts" (p. 168). This is not insurmountable but will certainly require innovative measurement approaches (e.g., Lee & von Davier, 2020; May, 2006; Rutkowski & Rutkowski, 2018). Thus, as others have done (Chudgar et al., 2014; Traynor & Raykov, 2013), we extend an invitation for further research on improving the reliability and comparability of family background measures. We believe the availability of reliable and comparable family background measures will aid efforts to investigate the HL Effect and other meaningful phenomenon cross-nationally.

Abbreviations

CSEM	Conditional standard error of measurement
HIV/AIDS	Human immunodeficiency virus/acquired immunodeficiency syndrome
HL Effect	Heyneman-Loxley Effect
ICC	Intraclass correlation coefficient
OLS	Ordinary least squares
PIRLS	Progress in International Reading Literacy Study
PISA	Program for International student assessment
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SES	Socioeconomic status
TIMSS	Trends in International Mathematics and Science Study

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-022-00139-3.

Additional file 1. Reading Naive and Correction Model Parameters for Botswana, Kenya, Lesotho, Malawi, and Mauritius. Reading Naive and Correction Model Parameters for Mozambique, Namibia, Seychelles, South Africa, and Eswatini. Reading Naive and Correction Model Parameters for Tanzania, Uganda, Zambia, Zanzibar, and Zimbabwe. Mathematics Naive and Correction Model Parameters for Botswana, Kenya, Lesotho, Malawi, and Mauritius. Mathematics Naive and Correction Model Parameters for Mozambique, Namibia, Seychelles, South Africa, and Eswatini. Mathematics Naive and Correction Model Parameters for Tanzania, Uganda, Zambia, Zanzibar, and Zimbabwe. HIV/AIDS Knowledge Naive and Correction Model Parameters for Botswana, Kenya, Lesotho, Malawi, and Mauritius. HIV/AIDS Knowledge Naive and Correction Model Parameters for Mozambique, Namibia, Seychelles, South Africa, and Eswatini. HIV/AIDS Knowledge Naive and Correction Model Parameters for Tanzania, Uganda, Zambia, Zanzibar, and Zimbabwe. National Income and Household Possessions Index Reliability in PISA 2009 Countries: Albania to Kyrgyzstan. National Income and Household Possessions Index Reliability in PISA 2009 Countries: Latvia to Uruguay. National Income and the Reliability of the Household Possessions Index in PISA 2009.

Acknowledgements
We wish to thank SACMEQ for providing generous access to the SACMEQ III data archive and Stephen Heyneman for his comments on an early draft of our manuscript.

Author contributions
Our contributions include the following: study conceptualization (WJR 40%, AA 30%, TFL 30%); literature review (WJR 60%, AA 20%, TFL 20%), analysis (WJR 100%); manuscript preparation (WJR 60%, AA 20%, TFL 20%); and manuscript revision (WJR 50%, AA 30%, TFL 20%). All authors read and approved the final manuscript.

Funding
We received no financial support for the research, authorship, or publication of the manuscript.

Availability of data and materials
SACMEQ III data archive is not publicly available. Researchers may request access at http://www.sacmeq.org/.

Declarations

Competing interests
We do not have competing interests.

Received: 25 May 2022 Accepted: 4 November 2022
Published online: 23 November 2022

References

Adelson, J. L., & McCoach, D. B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-type scale. *Educational and Psychological Measurement, 70*(5), 796–807. https://doi.org/10.1177/0013164410366694

Atteberry, A. C., & McEachin, A. J. (2020). Not where you start, but how much you grow: An addendum to the Coleman Report. *Educational Researcher, 49*(9), 678–685. https://doi.org/10.3102/0013189X20940304

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-Scale Assessments in Education, 8*, 1–37. https://doi.org/10.1186/s40536-020-00086-x

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. *Comparative Education Review, 46*(3), 291–312. https://doi.org/10.1086/341159

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

Bates, M. D., Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Handling correlations between covariates and random slopes in multilevel models. *Journal of Educational and Behavioral Statistics, 39*(6), 524–549. https://doi.org/10.3102/1076998614559420

Battauz, M., & Bellio, R. (2011). Structural modeling of measurement error in generalized linear models with Rasch measures as covariates. *Psychometrika, 76*(1), 40–56. https://doi.org/10.1007/S11336-010-9195-Z

Battauz, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics, 36*(3), 283–306. https://doi.org/10.3102/1076998610366262

Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics, 121*(4), 1437–1472. https://doi.org/10.1093/qje/121.4.1437

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality 1. *Bulletin of Sociological Methodology/bulletin De Méthodologie Sociologique, 66*(1), 60–75. https://doi.org/10.1177/075910630006600106

Borman, G., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201–1246. https://doi.org/10.1177/016146811011200507
Bos, K., & Kuiper, W. (1999). Modelling TIMSS data in a European comparative perspective: Exploring influencing factors on achievement in mathematics in grade 8. *Educational Research and Evaluation, 5*(2), 157–179. https://doi.org/10.1076/edre.5.2.157.6946
Bouhlila, D. S. (2015). The Heyneman-Loxley effect revisited in the Middle East and North Africa: Analysis using TIMSS 2007 database. *International Journal of Educational Development, 42*, 85–95. https://doi.org/10.1016/j.ijedudev.2015.02.014
Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Amsterdam: Elsevier.
Bowles, S., & Levin, H. M. (1968). The determinants of scholastic achievement-An appraisal of some recent evidence. *Journal of Human Resources, 3*(1), 3–24. https://doi.org/10.2307/144645
Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS*. Springer.
Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). National Academies Press.
Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Chapman & Hall/CRC Press.
Burkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01
Cain, G. G., & Watts, H. W. (1970). Problems in making policy inferences from the Coleman Report. *American Sociological Review, 35*(2), 228–242. https://doi.org/10.2307/2093201
Caro, D. H., & Cortes, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments, 5*, 9–33.
Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC Press.
Carver, R. P. (1975). The Coleman report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12*(1), 77–86. https://doi.org/10.3102/00028312012001077
Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39*(5), 333–367. https://doi.org/10.3102/1076998614547576
Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-, school-, and student-level analyses. *Journal of Family Psychology, 21*(3), 510–519. https://doi.org/10.1037/0893-3200.21.3.510
Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review, 84*(3), 517–544. https://doi.org/10.1177/0003122419847165
Chudgar, A., & Luschei, T. F. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Educational Research Journal, 46*(3), 626–658. https://doi.org/10.3102/0002831209340043
Chudgar, A., Luschei, T. F., & Fagioli, L. P. (2014). A call for consensus in the use of student socioeconomic status measures in cross-national research using the Trends in International Mathematics and Science Study (TIMSS). Teachers College Record. https://www.tcrecord.org/Content.asp?ContentId=17564
Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, D.C: U. S. Government Printing Office.
Cowan, C. D., Hauser, R. M., Kominski, R. A., Levin, H. M., Lucas, S. R., Morgan, S. L., Spencer, M. B., & Chapman, C. (2012). *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation (Recommendations for the National Center for Education Statistics)*. National Center for Education Statistics.
Cresswell, J., Schwantner, U., & Waters, C. (2015). *A review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. Washington DC, Paris: The World Bank and OECD Publishing. https://doi.org/10.1787/9789264248373-en
Culpepper, S. A. (2012). Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: The single-indicator case. *Applied Psychological Measurement, 36*(5), 349–374. https://doi.org/10.1177/0146621612446806
de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
Department of Education and Science. (1967). *Children and their primary schools: A report of the central advisory council for education (London)*. Her Majesty's Stationary Office and the Queen's Printer for Scotland.
Dolata, S. (2008). Indice du statut socioeconomique du milieu familial des eleves du SACMEQ: Construction avec le modele de Rasch et analyses. *Mesure Et Evaluation En Education, 31*(1), 121–149. https://doi.org/10.7202/1025011ar
Engzell, P. (2021). What do books in the home proxy for? A cautionary tale. *Sociological Methods & Research, 50*(4), 1487–1514. https://doi.org/10.1177/0049124119826143
Engzell, P., & Jonsson, J. O. (2015). Estimating social and ethnic inequality in school surveys: Biases from child misreporting and parent nonresponse. *European Sociological Review, 31*(3), 312–325. https://doi.org/10.1093/esr/jcv005
Fargas-Malet, M., McSherry, D., Larkin, E., & Robinson, C. (2010). Research with children: Methodological issues and innovative techniques. *Journal of Early Childhood Research, 8*(2), 175–192. https://doi.org/10.1177/1476718X09345412

Fowler, F., Jr., & Cosenza, C. (2009). Design and evaluation of survey questions. In L. Bickman & D. J. Rog (Eds.), *The SAGE handbook of applied social research methods* (pp. 375–412). Thousand Oaks, CA: SAGE Publications, Inc.
Fuchs, M. (2005). Children and adolescents as respondents. Experiments on question order, response order, scale effects and the effect of numeric values associated with response options. *Journal of Official Statistics, 21*(4), 701–725.
Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.
Gamoran, A., & Long, D. A. (2007). Equality of educational opportunity: A 40 year retrospective. In R. Teese, S. Lamb, M. Duru-Bellat, & S. Helme (Eds.), *International studies in educational inequality, theory, and policy* (pp. 23–47). Dordrecht: Springer.
Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136
Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling, 8*(3), 243–261. https://doi.org/10.1177/1471082X0800800302
Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66*(3), 361–396. https://doi.org/10.3102/00346543066003361
Gruijters, R. J., & Behrman, J. A. (2020). Learning inequality in Francophone Africa: School quality and the educational achievement of rich and poor children. *Sociology of Education, 93*(3), 256–276. https://doi.org/10.1177/0038040720919379
Gustafson, P. (2021). The impact of unacknowledged measurement error. In G. Y. Yi, A. Delaigle, & P. Gustafson (Eds.), *Handbook of measurement error models* (pp. 37–52). Chapman and Hall/CRC.
Hannum, E., Liu, R., & Alvarado-Urbina, A. (2017). Evolving approaches to the study of childhood poverty and education. *Comparative Education, 53*(1), 81–114. https://doi.org/10.1080/03050068.2017.1254955
Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141–164. https://doi.org/10.3102/01623737019002141
Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*(5), 481–502. https://doi.org/10.1016/S0272-7757(03)00038-4
Harris, D. N. (2007). Diminishing marginal returns and the production of education: An international analysis. *Education Economics, 15*(1), 31–53. https://doi.org/10.1080/09645290601133894
Harwell, M. (2019). Don't expect too much: The limited usefulness of common SES measures. *The Journal of Experimental Education, 87*(3), 353–366. https://doi.org/10.1080/00220973.2018.1465382
Harwell, M., Maeda, Y., Bishop, K., & Xie, A. (2017). The surprisingly modest relationship between SES and educational achievement. *The Journal of Experimental Education, 85*(2), 197–214. https://doi.org/10.1080/00220973.2015.1123668
Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives, 15*(4), 57–67. https://doi.org/10.1257/jep.15.4.57
Heyneman, S. P. (1989). Multilevel methods for analyzing school effects in developing countries. *Comparative Education Review, 33*(4), 498–504. https://doi.org/10.1086/446882
Heyneman, S. (2016). The Heyneman/Loxley effect: Three decades of debate. In S. McGrath & Q. Gu (Eds.), *Routledge handbook of international education and development* (pp. 150–167). Routledge.
Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high- and low-income countries. *American Journal of Sociology, 88*(6), 1162–1194. https://doi.org/10.1086/227799
Hox, J. J. (2008). Accommodating measurement errors. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *The international handbook of survey methodology* (pp. 387–402). New York, London: Erlbaum/Taylor & Francis.
Huang, F. L. (2010). The role of socioeconomic status and school quality in the Philippines: Revisiting the Heyneman-Loxley effect. *International Journal of Educational Development, 30*(3), 288–296. https://doi.org/10.1016/j.ijedudev.2009.10.001
Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Cappelle, F., Paviot, L., & Vellien, J. (2010). *SACMEQ III project results: Pupil achievement levels in reading and mathematics*. Paris: International Institute for Educational Planning (IIEP) and SACMEQ.
Ilie, S., & Lietz, P. (2010). School quality and student achievement in 21 European countries. The Heyneman-Loxley effect revisited. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 3*, 57–84.
Jerrim, J., & Micklewright, J. (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. https://doi.org/10.1093/esr/jcu072
Junker, B., Schofield, L. S., & Taylor, L. J. (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics, 1*(1), 1–19. https://doi.org/10.1186/2193-8997-1-4
Kim, S. W., Cho, H., & Kim, L. Y. (2019). Socioeconomic status and academic outcomes in developing countries: A meta-analysis. *Review of Educational Research, 89*(6), 875–916. https://doi.org/10.3102/0034654319877155
Konstantopoulos, S., & Borman, G. (2011). Family background and school effects on student achievement: A multilevel analysis of the Coleman data. *Teachers College Record, 113*(1), 97–132.
Kreuter, F., Eckman, S., Maaz, K., & Watermann, R. (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. https://doi.org/10.18148/srm/2010.v4i3.4283
Lee, J., & Borgonovi, F. (2022). Relationships between family socioeconomic status and mathematics achievement in OECD and Non-OECD countries. *Comparative Education Review, 66*(2), 199–227. https://doi.org/10.1086/718930
Lee, J., & Fish, R. M. (2010). International and interstate gaps in value-added math achievement: Multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education, 117*(1), 109–137. https://doi.org/10.1086/656348

Lee, J., Zhang, Y., & Stankov, L. (2019). Predictive validity of SES measures for student achievement. *Educational Assessment, 24*(4), 305–326. https://doi.org/10.1080/10627197.2019.1645590
Lee, S. S., & von Davier, M. (2020). Improving measurement properties of the PISA home possessions scale through partial invariance modeling. *Psychological Test and Assessment Modeling, 62*(1), 55–83.
Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22–52. https://doi.org/10.3102/1076998613509405
Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585. https://doi.org/10.1126/science.aal3618
Ludtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376
Lumley T. (2021). *Survey: Analysis of complex survey samples* (Version 4.1–1). https://cran.r-project.org/web/packages/survey/index.html
Marsh, H. W., Ludtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Koller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488
Marsh, H. W., Ludtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthen, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. https://doi.org/10.1080/00273170903333665
May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics, 31*(1), 63–79. https://doi.org/10.3102/10769986031001063
OECD. (2012). *PISA 2009 technical report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en
Pokropek, A., Borgonovi, F., & Jakubowski, M. (2015). Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences, 42*, 10–18. https://doi.org/10.1016/j.lindif.2015.07.011
R Development Core Team. (2021). R: A language and environment for statistical computing (Version 4.1.2). Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org. Accessed 3 Dec 2021.
Reynolds, D., & Creemers, B. (1990). School effectiveness and school improvement: A mission statement. *School Effectiveness and School Improvement, 1*(1), 1–3.
Riddell, A. R. (1989a). An alternative approach to the study of school effectiveness in third world countries. *Comparative Education Review, 33*(4), 481–497. https://doi.org/10.1086/446881
Riddell, A. R. (1989b). Response to Heyneman. *Comparative Education Review, 33*(4), 505–506. https://doi.org/10.1086/446883
Ridolfo, H., & Maitland, A. (2011). Factors that influence the accuracy of adolescent proxy reporting of parental characteristics: A research note. *Journal of Adolescence, 34*(1), 95–103. https://doi.org/10.1016/j.adolescence.2010.01.008
Rury, J., & Saatcioglu, A. (2015). Did the Coleman Report underestimate the effect of economic status on educational outcomes? *Teachers College Record*. Retrieved October 3, 2021, from http://www.tcrecord.org/content.asp?contentid=17828
Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278. https://doi.org/10.2304/rcie.2013.8.3.259
Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354–367. https://doi.org/10.1080/00313831.2016.1261044
Sandoval-Hernandez, A., Rutkowski, D., Matta, T., & Miranda, D. (2019). Back to thedrawing board: Can we compare socioeconomic background scales? *Revista De Educacion, 383*, 37–61. https://doi.org/10.4438/1988-592X-RE-2019-383-400
Schiller, K. S., Khmelkov, V. T., & Wang, X. Q. (2002). Economic development and the effects of family characteristics on mathematics achievement. *Journal of Marriage and Family, 64*(3), 730–742. https://doi.org/10.1111/j.1741-3737.2002.00730.x
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0
Simmons, J., & Alexander, L. (1978). The determinants of school achievement in developing countries: A review of the research. *Economic Development and Cultural Change, 26*(2), 341–357. https://doi.org/10.1086/451019
Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10.1037/pas0000648
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417
Stan Development Team (2021). *RStan: The R interface to Stan* (Version 2.21.1). http://mc-stan.org/
Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review, 67*(4), 639–652.
Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L. E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement, 26*(1), 75–101. https://doi.org/10.1080/09243453.2013.871302

Theisen, G. L., Achola, P. P., & Boakari, F. M. (1983). The underachievement of cross-national studies of achievement. *Comparative Education Review, 27*(1), 46–68. https://doi.org/10.1086/446345
Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research, 37*(5), 483–504. https://doi.org/10.1016/S0883-0355(03)00016-8
Traynor, A., & Raykov, T. (2013). Household possessions indices as wealth measures: A validity evaluation. *Comparative Education Review, 57*(4), 662–688. https://doi.org/10.1086/671423
Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956–972. https://doi.org/10.1177/0013164404268674
White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461. https://doi.org/10.1037/0033-2909.91.3.461
Woessmann, L. (2010). Families, schools and primary-school learning: Evidence for Argentina and Colombia in an international perspective. *Applied Economics, 42*(21), 2645–2665. https://doi.org/10.1080/00036840801964617
Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society: Series A (statistics in Society), 159*(2), 201–212. https://doi.org/10.2307/2983168
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
World Bank. (2009). *World development indicators 2009*. Washington DC: The World Bank.
World Bank. (2011). *World development indicators 2011*. Washington DC: The World Bank.
Yang, Y., & Gustafsson, J. E. (2004). Measuring socioeconomic status at individual and collective levels. *Educational Research and Evaluation, 10*(3), 259–288. https://doi.org/10.1076/edre.10.3.259.30268
Yi, G. Y. (2021). Likelihood methods with measurement error and misclassification. In G. Y. Yi, A. Delaigle, & P. Gustafson (Eds.), *Handbook of measurement error models* (pp. 99–125). Chapman and Hall/CRC.
Yi, G. Y., & Buzas, J. S. (2021). Measurement error models—A brief account of past developments and modern advancements. In G. Y. Yi, A. Delaigle, & P. Gustafson (Eds.), *Handbook of measurement error models* (pp. 3–36). Chapman and Hall/CRC.

Publisher's Note