# Critiques of socio-economic school compositional effects: Are they valid?

Michael G. Sciffer (Corresponding Author)
School of Education, Murdoch University, 90 South Street, Murdoch, Western Australia 6150.
ORCID: 0000-0003-2552-8840
research@sciffer.info

Laura B. Perry
Associate Professor Education Policy and Comparative Education, School of Education, Murdoch University.
ORCID: 0000-0003-4252-2379
l.perry@murdoch.edu.au

Andrew McConney
Associate Professor Program Evaluation, Research Methods and Classroom Assessment, School of Education, Murdoch University.
ORCID: 0000-0001-7618-8829

***This is an original manuscript/preprint of an article published by Taylor & Francis in the British Journal of Sociology of Education on 24/4/2020, available online: http://www.tandfonline.com/10.1080/01425692.2020.1736000***

***The published article in the British Journal of Sociology of Education contains substantial modifications.***

**Abstract**

Recent studies have rejected school socio-economic compositional effects based on criticisms of the methodologies of prior studies and their own findings. We respond to the critiques of ecological fallacies and lack of control for prior achievement in school compositional research. We describe how prior ability control variables and fixed-effects methods have been inappropriately applied in research critical of compositional effects. We demonstrate that structural equation modeling can address concerns about the inflation of level-2 effects due to level-1 measurement error whilst also finding significant socio-economic school compositional effects. We conclude that the veracity of school socio-economic composition effects has not been weakened by recent critical studies and remain a profound issue for researchers and policymakers.

The socio-economic compositional (SEC) effect is the relationship between the socio-economic profile of a school and individual student outcomes such as academic achievement, attendance, completion and tertiary entrance (Rumberger & Palardy, 2005). It is a separate construct from the relationship between individual student socio-economic status (SES) and performance outcomes (Rumberger & Palardy, 2005). It can be thought of as the difference in performance between two students who have the same SES due to attending schools with different socio-economic profiles (Raudenbush and Bryk, 2002, p. 141). SEC is usually measured through the aggregation of the SES of the students in a school or class (Willms, 2010).

A range of mechanisms have been identified that explain the relationship between SEC and schooling outcomes. Rumberger and Palardy (2005) found that teacher expectations, hours of homework, number of academic courses taken, and student's sense of safety mediate the relationship between SEC and academic achievement growth. Willms (2010) found that quality of instruction, student engagement, curriculum coverage, instructional time, and adequacy of school resources mediate SEC and academic achievement. Palardy (2013) found that peer effects and school resources mediate SEC and graduation and college enrolment rates.

School composition has been an influential construct in school effectiveness research and broader policy reforms since the Coleman Report found that it accounted for between 3% and 33% of the variation in performance between schools, depending on ethnic background and grade (Coleman et al., 1966, p. 299). Lamb and Fullarton (2002) found that socio-economic composition and tracking had larger effects at the classroom and school levels on mathematics achievement than teacher quality, student attitudes, student beliefs and amount of homework in Australia and the US. Chiu and Khoo (2005) found the degree of clustering of students in schools according to parental occupational status inversely related to national-level academic performance in mathematics and science. In his meta-analytical review of the literature, Sirin (2005) found that aggregated measures of SES were stronger predictors of student outcomes than student-level measures[1]. Successive cycles of the OECD's Programme for International Student Assessment (PISA) have consistently found that socio-economic factors are stronger predictors of academic achievement between schools compared to within schools in most participating countries (OECD, 2003, 2004, 2007, 2010, 2013, 2016). Policy responses have led to programmes aimed at reducing socio-economic segregation between schools (Kahlenberg, 2007) and targeting resources to lower SEC schools (Gonski et al., 2011).

However despite this substantial body of research evidence, a series of studies by Marks and colleagues (Armor, Marks & Malatinsky, 2018; Marks, 2010, 2015, 2017) have challenged the substantiveness of SEC effects based on their view of research methodologies used in school compositional research and their own research findings. They have argued that school compositional effects may be statistical artefacts arising from inappropriate methodologies. These works are a subset of Marks' broader research programme critical of the role of socio-economic status in education policy (Marks, 2014, 2016, 2017). Marks has argued that education policymakers and researchers have had an unwarranted focus on SES given that genetic and cognitive differences, not SES, are the dominant causes of diversity of student outcomes (Marks, 2017).

This article will respond to Marks and colleagues' (Armor et al., 2018; Marks, 2010, 2015, 2017) critiques of the veracity of SEC effects in three sections. Firstly, we will address their arguments that prior school compositional research has suffered from ecological fallacies and lack of control of prior achievement. Secondly, we will consider Marks and colleagues' application of residualised change and fixed effects analyses that have found null SEC effects. Finally, we will demonstrate that structural equation modeling (SEM) can be used to address Marks and colleagues' concerns about measurement error inflating SEC effects in multilevel research.

## Critiques of Prior School Compositional Research

Marks and colleagues (Armor et al., 2018; Marks, 2010, 2015, 2017) have critiqued school compositional research for ecological fallacies and lack of control for prior achievement but have rarely engaged with specific studies. The ecological fallacy is when a relationship observed at the group level is mistakenly used to describe the effect of group membership on individuals (Robinson, 1950). Marks and colleagues have only referred to Hauser's (1970) critique of cross-tabulation methods and White, Reynolds, Thomas and Gitzlaff's (1993) critique of aggregated measures in single-level regressions when arguing that school compositional research suffers from the ecological fallacy. This ignores the large body of research that has utilized multi-level modeling (MLM) techniques to find SEC effects (Ewijk & Sleegers, 2010). MLM addresses the ecological fallacy through the aggregation of individual-level parameters to the group level (Snijders & Bosker, 2012, p. 83). Thus, MLM avoids the ecological fallacy as compositional effects are not the relationship between a group-level predictor and the dependent variable, but the relationship between the difference of an aggregated group-level predictor and its associated individual-level predictor, and the dependent variable (Raudenbush & Bryk, 2002, pp. 139-141).

Marks and colleagues (Armor et al., 2018; Marks 2015, 2017) have also argued that many school compositional studies have overestimated SEC effects by not controlling for prior achievement. This criticism has not considered the differing aims of longitudinal and cross-sectional compositional research. Compositional research that controls for prior achievement can estimate compositional effects over specific time periods in a school career. For example, Rumberger and Palardy (2005) examined the effect of SEC on achievement growth from grades 8 to 12 in US high schools. On the other hand, cross-sectional studies estimate the total effect associated with SEC up to the time point of a single academic measurement. For example, Willms (2010) internationally compared the effects of SEC on 15-year-old's scientific literacy. Typically, larger compositional effect sizes in cross-sectional models can be accounted for by SEC being accumulative, a common assumption in school effectiveness research (Sass, Semykina & Harris, 2014). It is not appropriate to dismiss such models as they allow researchers and policymakers to evaluate systemic differences in compositional effects (Willms, 2010).

## Marks and Colleagues' Methodologies

Marks and colleagues have based much of their critiques of school compositional effects on the findings of their studies. As we show in this section, however, their studies contain several methodological flaws, which undermine the persuasiveness of their arguments.

### Residualised Change Models

A critical methodological flaw in Marks' (2010, 2015) and Armor, Marks & Malatinsky's (2018) residualised change models is the methodological misapplication of prior achievement. Residualised change models are two-occasion growth models that control for prior levels of a dependent variable by including it as a covariate in regressions (Gollwitzer, Christ & Lemmer, 2014). Prior achievement is included in residualised change models to create "quasi-gain" models (Schochet & Chiang, 2010) to allow an estimate of the effects of other predictor variables whilst controlling for the effect of prior achievement (Castro-Schilo & Grimm, 2018). Such models remove all of the effect of prior inputs and processes that are associated with academic performance, such as individual ability, school resources, SES, SEC, parental engagement and teaching practices. They allow for the measurement of the "value add" of schools, teachers or interventions by controlling for learning prior to the first-occasion measurement of achievement (Raudenbush & Willms, 1995).

Marks (2010, 2015) and Armor et al. (2018) have extended the purpose of prior achievement in residualised change models to also serve as a comparative predictor of academic growth alongside SEC and found that it had a larger effect on academic growth than SEC. This is unsurprising as achievement and prior achievement are measuring the same latent construct – academic achievement. Problematically they use this finding to conclude that SEC is thus inconsequential. This interpretation of the comparative effect size of prior achievement would result in no predictor variables being of substantive interest to school effectiveness researchers in residualised change models. It is analogous to arguing that because height at age 13 is the strongest predictor of height at age 15, then diet is of no substantive interest.

**Fixed Effects Analyses**

Claims by Marks (2015) and Armor et al. (2018) that their fixed effects models demonstrated that SEC effects do not exist are unfounded as fixed effects analyses are incapable of modeling compositional effects. Fixed effects models control for unobserved differences between participants in multiple-occasion data by removing all of the between-individual time-invariant differences from the analysis (Allison, 2011, pp. 2-4). Thus, fixed-effects models only provide estimates of within-individual effects that change with time. They are unable to estimate school compositional effects like SEC, because it is a between-student effect (Bell, Fairbrother & Jones, 2018). Marks (2015) and Armor et al. (2018) nevertheless attempted to measure SEC with fixed-effects analyses by operationalising it as a within-individual effect. As such, their models are only able to estimate the effects of changes in SEC on academic achievement growth. Such changes are likely negligible from year-to-year within the same school cohorts. Thus their conclusions that SEC has little effect on academic achievement growth were largely due to the limitations of their methodology.

For example, *Equation 1* is a two-occasion fixed effects model of the difference in achievement scores, or achievement growth, due to changes in SEC:

$$Y_{i2} - Y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \tag{1}$$

where $Y_{i2} - Y_{i1}$ is the difference in academic achievement of student $i$ between the two time periods, $(\mu_2 - \mu_1)$ is the difference between intercepts, $\beta$ is the coefficient for the difference in SEC scores, and $(\varepsilon_{i2} - \varepsilon_{i1})$ is the difference in individual error between the two time periods. If $x_{i2} \approx x_{i1}$, that is, SEC negligibly changes, then it is unlikely that a statistically significant effect for changes in SEC would be detected by a fixed effects analysis as $\beta$ will be close to zero.

Marks' (2015) fixed effects analyses were five 3-occasion models of academic achievement in Australia's National Assessment Program – Literacy and Numeracy (NAPLAN) which attempted to detect the effect of changes in SEC on changes in academic achievement. The dependent variables were academic difference scores in numeracy, reading, writing, spelling and grammar for students in Years 3, 5 and 7 in Victorian state schools. The study design was unlikely to capture SEC change effects as the dependent variable spanned primary school where SEC changes are negligible, and it inappropriately utilised a measure of high school SEC to explain primary school academic achievement growth.

Years 3 and 5 are primary school grades and Year 7 is the first year of high school. Over 90% of the time period between the Year 5 and Year 7 assessments is primary schooling, thus the Year 7 tests are largely a measure of primary school learning (Lu & Rickard, 2014). Evidence also suggests that few primary school students change school each year in Australia (Lu & Rickard, 2016) resulting in little change in the SEC of a primary school cohort over two years. Thus, Marks' 2015 study was very unlikely to detect an effect of changes in SEC on changes in academic achievement as SEC change scores would be close to zero. Additionally, the finding of small negative estimates for changes in SEC may be explained by the negative effects of school change during primary school (Reynolds, Chen & Herbers, 2009) counteracting the positive effects of children moving to higher-SEC schools.

The second weakness in Marks' (2015) design was to explain the academic difference between Years 5 and 7 with the change in SEC from Year 5 to 7. As previously mentioned, Year 7 NAPLAN tests are largely a measure of primary school achievement. Therefore, very little of the difference between Year 5 and 7 would be explained by the measure of change in school context (SEC) from primary to high school.

The fixed effects analyses in Armor et al. (2018) were separate 6-occasion models of math and reading achievement in Grades 3 to 8 in North Carolina, South Carolina and Arkansas. The models attempted to detect the effect of changes in SEC on changes in academic achievement. Unlike Marks (2015), this design does capture a change in school context from elementary to middle school. But this SEC change effect was unlikely to be detected as it was attenuated by including 5 occasions without cohort changes in school context in the same model. These other occasions with minimal SEC changes were averaged against the change in SEC from elementary to middle school, diminishing its effect. As well, the high proportion of public school attendance in the US suggests there may be minimal SEC changes from elementary to middle schools. Similar to Marks (2015), student mobility within elementary and middle school grades also likely confounded positive SEC effects. Thus, like Marks (2015), this study's design was unlikely to detect an effect of changes in SEC on academic achievement growth.

## Measurement and Aggregation Error in School Compositional Research

Marks and colleagues' (Armor et al., 2018; Marks, 2010, 2015, 2017) have also highlighted the issue of false or "phantom" compositional effects due to potential error in measures of SES (Pokropek, 2015) inflating SEC effects in multilevel models. Simulation research by Pokropek (2015) demonstrated that increases in the unreliability of level-1 variables inflate the effect sizes of level-2 variables that are aggregates of level-1 variables. Simulation (Pokropek, 2015) and applied studies (Pokropek, 2015; Televantou et al., 2015) have shown that structural equation modeling can appropriately addresses level-1 measurement error to reliably estimate aggregated level-2 effects. Notably, Marks' (2010, 2015) own aggregations of prior achievement have also been shown (Dumay & Dupriez, 2008) to be potentially subject to the same measurement error as other compositional effects.

## Empirical Demonstration

We conducted a methodological exploration of a subsample of the Programme for International Student Assessment's (PISA) publicly available 2015 dataset to demonstrate a means to address the potential for level-1 measurement error inflating compositional effects. No attempt is made to generalise the findings of this demonstration other than to show SEC effects can be detected after controlling for level-1 measurement error.

PISA is an international triennial assessment of the academic achievement of 15-year-olds in reading, mathematics and science (OECD, 2016). Reading achievement was the dependent variable in our models. We selected five subsamples of the dataset (Australia, Brazil, Germany, Indonesia, Japan and the US) as a diverse set of schooling systems, of which Australia and the US were previously analysed by Marks and colleagues' (Armor et al., 2018; Marks, 2010, 2015, 2017). These samples have average population coverages at the school level ranging from 19% to 70%.

PISA's measure of SES, the index of economic, social and cultural status (ESCS), is derived from the first component of a principal components analysis (PCA) of student-reported sub-indices of home possessions (HOMEPOS), highest parental occupation (HISEI) and highest parental education (PARED) (OECD, 2017, p. 339). Measures based on PCA can contain measurement error (Dunteman, 1989, p. 60), thus measures of SEC aggregated from ESCS may be inflated.

We first constructed a hierarchical regression model (HRM) to obtain coefficient estimates of SES and SEC for each subsample. We used country-specific measures of SES derived from the first component of a PCA of the subindices of ESCS instead of the OECD's internationally-derived ESCS. This was to allow a direct comparison with our SEMs which also had country-specific factor loadings for SES and SEC[2].

The HRM is represented in *Equation 2* (Raudenbush & Bryk, 2002, p. 140) where $\beta_1$ is the coefficient for the within-school effect of SES and $\beta_2$ is the coefficient for the averaged between-school effect of SES. The coefficient for SEC was derived from *Equation 3* (Raudenbush & Bryk, 2002, p. 139).

$$Y_{ij} = \beta_0 + \beta_1\left(X_{ij} - \bar{X}_{.j}\right) + \beta_2\bar{X}_{.j} + \delta_{0j} + \varepsilon_{ij} \qquad (2)$$

$$\beta_c = \beta_2 - \beta_1 \qquad (3)$$

Secondly, we developed latent-manifest (L-M) and latent-latent (L-L) SEMs analogous to equation 2 (Marsh et al., 2009) where within- and between-school effects of SES were operationalised through the sub-indices of ESCS. The L-M SEM is represented in *Figure 1* and the L-L SEM in *Figure 2*. Again, the coefficient for SEC was derived from equation 3. We followed Marsh and colleagues' (2009) advice comparing L-M and L-L SEMs as L-M SEMs may underestimate the size of compositional effects with formative constructs such as SES when clusters are small samples (Grilli & Rampichini, 2011).

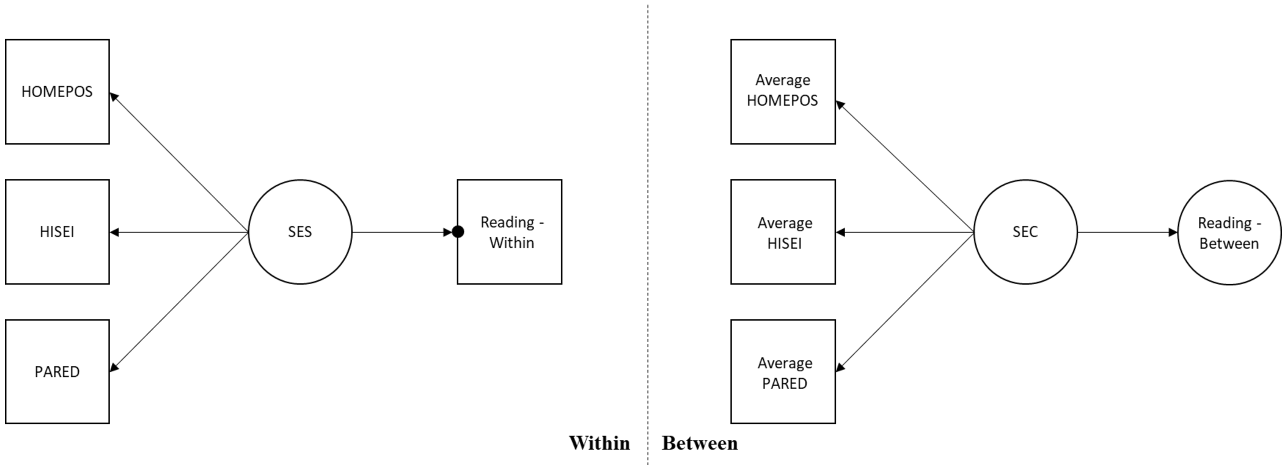


*Figure 1.* Latent-manifest structural equation model of SES and SEC on reading.

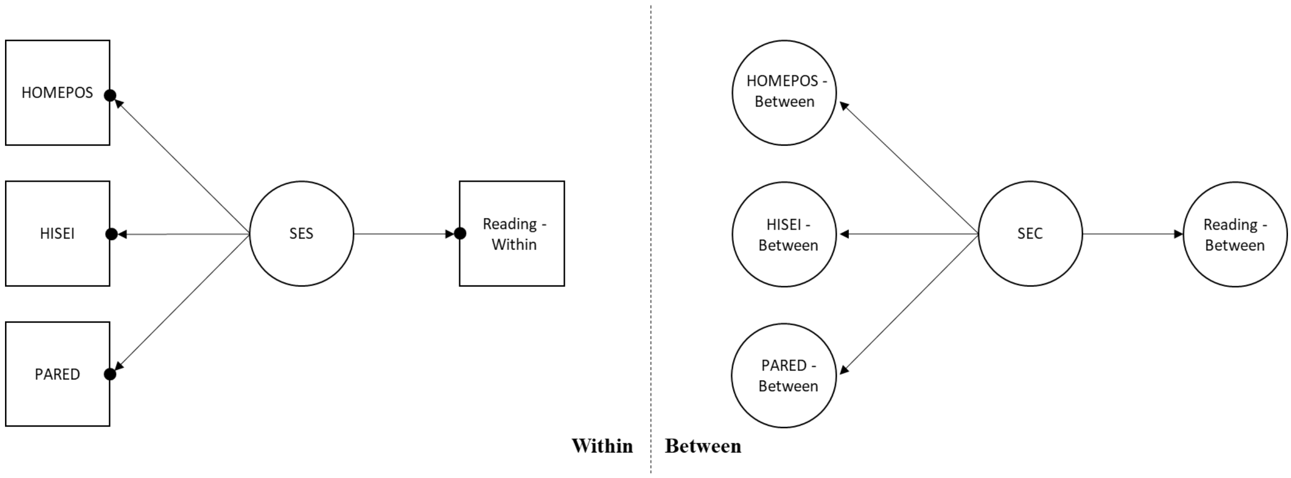


*Figure 2.* Latent-latent structural equation model of SES and SEC on reading.

Formative constructs are aggregates derived from individual measures of level-1 constructs, such as SES (Lüdtke et al., 2008). Differences between the level-1 scores that underlie formative constructs consist of differences between individuals and measurement error, thus individual scores are not interchangeable as individuals rate their own characteristics. Alternatively, reflective constructs are aggregates derived from individual measures of level-2 constructs, such as school climate (Lüdtke et al., 2008). Differences between scores only consist of measurement error, thus individual scores are interchangeable as individuals are rating a shared environment. The operationalisation of formative and reflective constructs is discussed in Grilli and Rampichini (2011), Lüdtke et al. (2008), and Marsh et al. (2009).

PCA was performed with the *prcomp()* package in *R 3.5.3* (R Core Team, 2019). Missing data in the PCA was handled through a single bootstrapped-imputation with the *Amelia II* (Honaker, King & Blackwell, 2011) package. HRMs and SEMs were performed with *Mplus 8.2* (Muthén & Muthén, 2017). Missing data in HRMs and SEMs were handled with the full information maximum-likelihood method.

## Findings and Analysis

*Table 1* summarises the findings of the three modeling approaches. In all national samples HRM found slightly smaller standardised coefficients for SES than the SEMs of between 0.01 to 0.06 standard deviations lower. This suggests that level-1 measurement error may minimally attenuate SES effects in HRMs with PCA-based measures in many national samples.

Table 1

*Hierarchical Regression and Structural Equation Models of SEC on Reading in PISA 2015 Subsamples*

| Country/(*M* C)[a] | | HRM | L-M SEM | L-L SEM |
|---|---|---|---|---|
| Australia/(19.25%) | | | | |
| | Intercept | 493.973 | 492.897 | 493.400 |
| | Intercept variance | 984.750 | 829.534 | 788.131 |
| | Residual variance | 8068.138 | 7865.448 | 7841.801 |
| | $\beta_1$ | 0.153 | 0.206 | 0.213 |
| | $\beta_2$ | 0.374 | 0.394 | 0.386 |
| | $\beta_c$ | 0.266 | 0.227 | 0.227 |
| | RMSEA | | 0.016 | 0.015 |
| | CFI | | 0.990 | 0.994 |
| Brazil/(55.53%) | | | | |
| | Intercept | 388. 368 | 388.105 | 388.001 |
| | Intercept variance | 2260.518 | 1935.766 | 1802.680 |
| | Residual variance | 5969.936 | 5951.059 | 5945.646 |
| | $\beta_1$ | 0.035* | 0.054 | 0.054 |
| | $\beta_2$ | 0.490 | 0.519 | 0.530 |
| | $\beta_c$ | 0.457 | 0.457 | 0.474 |
| | RMSEA | | 0.025 | 0.026 |
| | CFI | | 0.962 | 0.980 |

Table 1 (continued)

| | | | | |
|---|---|---|---|---|
| Germany/(45.12%) | | | | |
| | Intercept | 477.868 | 477.595 | 478.060 |
| | Intercept variance | 2181.782 | 1191.408* | 799.287 |
| | Residual variance | 5436.826 | 5397.555 | 5389.706 |
| | $\beta_1$ | 0.065 | 0.087 | 0.088 |
| | $\beta_2$ | 0.557 | 0.632 | 0.659 |
| | $\beta_c$ | 0.512 | 0.561 | 0.590 |
| | RMSEA | | 0.038 | 0.046 |
| | CFI | | 0.950 | 0.958 |
| Indonesia/(69.75%) | | | | |
| | Intercept | 380.754 | 380.738 | 380.908 |
| | Intercept variance | 1027.929 | 881.741 | 864.072 |
| | Residual variance | 3452.317 | 3444.304 | 3443.294 |
| | $\beta_1$ | 0.040** | 0.054** | 0.054** |
| | $\beta_2$ | 0.408 | 0.440 | 0.444 |
| | $\beta_c$ | 0.372 | 0.387 | 0.395 |
| | RMSEA | | 0.012 | 0.008 |
| | CFI | | 0.995 | 0.998 |
| Japan/(20.76%) | | | | |
| | Intercept | 499.026 | 499.016 | 499.022 |
| | Intercept variance | 1594.460 | 1380.142 | 1182.834 |
| | Residual variance | 5101.190 | 5017.113 | 5012.266 |
| | $\beta_1$ | 0.058* | 0.110* | 0.112* |
| | $\beta_2$ | 0.537 | 0.558 | 0.577 |
| | $\beta_c$ | 0.503 | 0.461 | 0.479 |
| | RMSEA | | 0.041 | 0.046 |
| | CFI | | 0.900 | 0.918 |
| US/(33.04%) | | | | |
| | Intercept | 490.257 | 491.840 | 490.790 |
| | Intercept variance | 1713.695 * | 1921.392* | 1581.742 |
| | Residual variance | 7309.937 | 7254.924 | 7228.813 |
| | $\beta_1$ | 0.129 | 0.160 | 0.162 |
| | $\beta_2$ | 0.320 | 0.253* | 0.324 |
| | $\beta_c$ | 0.227* | $0.150^{ns}$ | 0.210* |
| | RMSEA | | 0.028 | 0.018 |
| | CFI | | 0.941 | 0.991 |

*Note.* All $p < .001$ unless $*p < .01$ or $**p < .05$ or $^{ns}$. All coefficients were standardised on the total model variance to make them comparable. Separate analyses were run for each of the 10 plausible values for reading and combined by Rubin's rules (Rubin, 1987). Both student- and school-level weights were applied.

Table 1 (continued)
[a]*M* C is the average school-level coverage of 15-year-olds.

In the Australian, Japanese and US samples HRM found larger coefficients for SEC than SEMs. In the other national samples, the HRM found smaller coefficients for SEC than the SEMs, apart from Brazil where the coefficient for SEC in the HRM was equal to the L-M SEM. This suggests that in studies where clusters sample a low proportion of the target population, level-1 measurement error may inflate aggregated level-2 effects in HRMs with PCA-based measures. In other cases, HRMs may underestimate compositional effects.

In each sample, L-L SEMs found larger SEC coefficients than L-M SEMs except for Australia where the coefficients were equal. Lüdtke et al. (2008) found that latent aggregation may be most appropriate with formative constructs with low population coverage. In this study, the samples with population coverages below 50% differed in terms of the RMSEA model fit index and the comparative sizes of within- and between-school effects. In Germany and Japan, RMSEA showed poorer fit in L-L compared to L-M SEMs and between-school SEC coefficients were more than 4 times the size of within-school SES coefficients. In Australia and the US, RMSEA showed better fit in L-L compared to L-M SEMs and between-school SEC coefficients were less than twice the size of within-school SES coefficients. Thus, it may be more appropriate to use latently aggregated SEMs with formative constructs when population coverage is low and compositional effects are of a similar magnitude to within-group effects. In the cases of Australia and the US, comparing modeling approaches suggests HRM may have overestimated the SEC effect by 0.04 and 0.02 standard deviations respectively.

Overall it can be seen that hierarchical models of latent measures of SES and aggregated SEC can estimate models free from level-1 measurement error and the potentially attendant inflation of school compositional effects. The models in this demonstration show that SEC effects remain statistically significant after handling level-1 measurement error. The national samples examined in this study found that the effect of measurement error was small for both level-1 and level-2 effects being less than 0.1 standard deviations. The demonstration also suggests that it cannot be assumed that prior compositional research that has not controlled for level-1 measurement error is discredited as the potential upward bias is likely to be very small. In some cases, hierarchical regression models may have underestimated compositional effects. This demonstration is also consistent with the advice by Marsh et al. (2009) that it is beneficial to compare manifest and latent aggregation methods when sampling ratios approach 100%.

The benefits of utilising SEMs over HRMs when measuring SEC effects are threefold. Firstly, by addressing measurement error they increase confidence in the statistical significance of compositional effects. Secondly, they provide a more accurate estimation of the comparative sizes of SES and SEC effects. Thirdly, they may increase the power of detecting mediating factors between SES and academic achievement with large samples.

A limitation of the use of latent modeling to address measurement error in SES variables is that single-factor models require three indicators to be identified (Bollen, 1989, p. 244). Not every educational dataset meets this requirement. For example, some US research can only measure SES by student access to free and reduced-price lunch eligibility (Domina et al., 2018). It is possible to measure a SES factor with two-indicators, but such models require other factors to be identifiable (Bollen, 1989).

## Conclusion

Marks' and colleagues' (Armor et al., 2018; Marks, 2010, 2015, 2017) criticisms of school compositional research have not weakened the veracity of SEC effects. Contemporary school compositional research is typically not marred by the ecological fallacy and the potential for this is accounted for by multilevel modeling. Cross-sectional studies are a valid methodology for evaluating systemic differences between education systems. The misapplication of prior achievement in residual change models explains the finding of comparatively small SEC effect sizes. The incapacity of fixed-effects analyses to measure compositional effects explains null findings for SEC. The issue of measurement error inflating compositional effects is a valid methodological criticism, but our findings show that it may be a small bias of less than 0.05 standard deviations in PISA samples and can be addressed through SEM.

School compositional effects are a profound issue for researchers and policymakers. Primary analysis of the most recent PISA (OECD, 2016, pp. 225-227) showed that across all participating countries, schools accounted for 87% of the effect of socio-economic factors on academic achievement. This suggests that many schooling systems are compounding the detrimental effects of social disadvantage on children's learning through school structures and policies that concentrate disadvantaged students into disadvantaged schools. Future school compositional research would benefit from expanding to non-academic performance measures and exploring the factors that mediate SEC relationships with schooling outcomes. This may indicate new options for equity-based school policy reforms to address the potentially broader implications of school SEC beyond academic achievement.

## References


Allison, P. D. (2011). *Fixed effects regression models.* Thousand Oaks, CA: Sage.

Armor, D. J., Marks, G. N., & Malatinszky, A. (2018). The impact of school SES on student achievement: Evidence from US statewide achievement data. *Educational Evaluation and Policy Analysis*, *40*(4), 613–630.

Bell, A., Fairbrother, M., & Jones, K. (2018). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*(2), 1051-1074.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, *35*(1), 32-58.

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: Country, school, and student level analyses. *American Educational Research Journal*, *42*(4), 575-603.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity* (Report No. OE-38001). Washington, DC: Government Printing Office.

Domina, T., Pharris-Ciurej, N., Penner, A. M., Penner, E. K., Brummet, Q., Porter, S. R., & Sanabria, T. (2018). Is free and reduced-price lunch a valid measure of educational disadvantage? *Educational Researcher*, *47*(9), 539–555.

Dumay, X., & Dupriez, V. (2008). Does the school composition effect matter? Evidence from Belgian data. *British Journal of Educational Studies, 56*(4), 440–477.

Dunteman, G. H. (1989). *Principal components analysis.* Newbury Park, CA: Sage.

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology, 44*(7), 673-682.

Gonski, D., Boston, K., Greiner, K., Lawrence, C., Scales, B., & Tannock, P. (2011). *Review of funding for schooling: Final report.* Retrieved from https://docs.education.gov.au/system/files/doc/other/review-of-funding-for-schooling-final-report-dec-2011.pdf

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*(4), 121.

Hauser, R. (1970). Context and consex: A cautionary tale. *American Journal of Sociology, 75*(4), 645-664.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1-47.

Kahlenberg, R. D., (2007). Socioeconomic school integration. *North Carolina Law Review, 85*(5), 1545–1594.

Lamb, S., & Fullarton, S. (2002). Classroom and school factors affecting mathematics achievement: A comparative study of Australia and the United States using TIMSS. *Australian Journal of Education, 46*(2), 154-171.

Lu, L., & Rickard, K. (2014). *Value added models for NSW government schools.* Retrieved from https://www.cese.nsw.gov.au//images/stories/PDF/VAPaper_v3-1Final.pdf

Lu, L., & Rickard, K. (2016). *Mobility of students in NSW government schools.* Retrieved from https://www.cese.nsw.gov.au/images/stories/PDF/student_mobility_report_feb16_AA.pdf

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203-229.

Marks, G. N. (2010). What aspects of schooling are important? School effects on tertiary entrance performance. *School Effectiveness and School Improvement, 21*(3), 267–287.

Marks, G. N. (2014). Demographic and socioeconomic inequalities in student achievement over the school career. *Australian Journal of Education*, *58*(3), 223-247.

Marks, G. N. (2015). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. *Oxford Review of Education, 41*(1), 122–144.

Marks, G. N. (2016). The relative effects of socio-economic, demographic, non-cognitive and cognitive influences on student achievement in Australia. *Learning and Individual Differences, 49*, 1-10.

Marks, G. N. (2017). Is SES really that important for educational outcomes in Australia? A review and some recent evidence. *The Australian Educational Researcher, 44*(2), 191–211.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, 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.

Muthén, L. K., &  Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Retrieved from http://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Organisation for Economic Co-operation and Development. (2003). *Literacy skills for the world of tomorrow: Further results from Pisa 2000*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development. (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1: Analysis.* Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development. (2010). *PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes (Volume II).* Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development (2013). *PISA 2012 results: Excellence through equity (Volume II): Giving every student the chance to succeed*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development (2017). *PISA 2015 technical report.* Paris, France: OECD Publishing.

Palardy, G. J. (2013). High school socioeconomic segregation and student attainment. *American Educational Research Journal*, *50*(4), 714-754.

Pokropek, A. (2015). Phantom effects in multilevel compositional analysis: Problems and solutions. *Sociological Methods & Research, 44*(4), 677–705.

R Core Team (2019). R: A language and environment for statistical computing. [Computer software]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Reynolds, A. J., Chen, C.-C., & Herbers, J. E. (2009, June). School mobility and educational success: A research synthesis and evidence on prevention. Paper presented at the Workshop on the Impact of Mobility and Change on the Lives of Young Children, Schools, and Neighborhoods, Washington, DC.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: John Wiley & Sons.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record, 107*(9), 1999.

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review, 38,* 9–23.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE Report 2010-4004). Retrieved from https://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* London: Sage.

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.

Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*(2), 134-150.

White, S. B., Reynolds, P. D., Marc Thomas, M., & Gitzlaff, N. J. (1993). Socioeconomic status and achievement revisited. *Urban Education, 28*(3), 328–343.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record*, *112*(4), 1008–1037.

## Endnotes

1. Sirin cautioned that the effect size for aggregated measures of SES in his meta-analysis may have been biased by the ecological fallacy.

2. We found that a two-level factor analysis of the subindices of ESCS, where holding factors equal across all countries in the full international sample, did not fit the data, having insignificant factor loadings. This is consistent with the OECD's report (OECD, 2017, p. 340) of differential national item loadings on ESCS. A drawback of allowing item loadings to vary by country is that factors are not internationally comparable.