Leading education
and social research
Institute of Education
University of London

Department of Quantitative Social Science

# The impact of sampling variation on peer measures: a comment on a proposal to adjust estimates for measurement error

Pedro N. Silva
John Micklewright
Sylke V. Schnepf

DoQSS Working Paper No. 12-12
December 2012

## Disclaimer

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science, Institute of Education, University of London
20 Bedford Way, London WC1H 0AL, UK

# The impact of sampling variation on peer measures: a comment on a proposal to adjust estimates for measurement error

**Perdo N. Silva[1], John Micklewright[2] and Sylke V. Schnepf[3]**

## Abstract

Investigation of peer effects on pupil's achievement with survey data on samples of schools and pupils within schools may mean that only a random sample of peers is observed for each individual pupil. This generates classical measurement error on peer variables. Hence under OLS model fitting the estimated peer group effects in a regression model are biased towards zero (attenuation). A simple adjustment for this kind of measurement error was proposed by Neidell and Waldfogel (2008). We review the derivation of the simple adjustment and suggest that it is not properly justified.

**JEL classification:** C21, C81, I21.

**Keywords:** Peer effects, measurement error, school surveys, sampling variation.

[1] Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro (pedro-luis.silva@ibge.gov.br)

[2] Department of Quantitative Social Science, Institute of Education, University of London (j.micklewright@ioe.ac.uk)

[3] Southampton Statistical Sciences Research Institute and School of Social Sciences, University of Southampton (svs@soton.ac.uk)

When peer effects are estimated in regression models with survey data that contain only a sample of each individual's peers, the estimates which are obtained can be expected to be biased. If the sample of peers is drawn randomly, as often occurs in surveys of school children, for example, the survey design leads to measurement error in the peer variables that is close (but not identical) to the classical textbook form, leading to attenuation bias. The problem has been recognised by Ammermüller and Pischke (2006, 2009) and subsequently by Waldfogel and Neidell (2008, 2010).[4]

In this note we consider these authors' proposal for a simple adjustment to the OLS estimator of the peer effects coefficient. We limit ourselves to discussing the presentation in Neidell and Waldfogel (2008) (NW from now on); see their equation (2) and their appendix which focuses on this issue. Their proposed adjustment follows work by Ammermueller and Pischke (2006, 2009) who consider in addition various other sources of measurement error which we are not concerned with here.

The original model considered by NW is given by:

$$y_{icd}^{t} = \beta_0 w_{icd} + \beta_1 \bar{W}_{icd} + \boldsymbol{\beta}_2' \mathbf{x}_{icd} + \boldsymbol{\beta}_3' \bar{\mathbf{X}}_{icd} + \boldsymbol{\beta}_4' \mathbf{z}_{cd} + \beta_5 y_{icd}^{t-1} + \alpha_d + \eta_{cd} + \varepsilon_{icd} \qquad (1)$$

where $y_{icd}^{t}$ denotes the outcome in kindergarten for child *i* in class *c* of school *d*;

$w_{icd}$ is the indicator that the child was enrolled in pre-school;

$\bar{W}_{icd} = (N_{cd} - 1)^{-1} \sum_{j \in U_{cd} - i} w_{jcd}$ is the average of the enrolment indicator for the population of peers in the class (i.e. excluding child *i*);

$U_{cd}$ is the set of children and $N_{cd}$ is the number of children in the class;

$\mathbf{x}_{icd}$ is a vector of individual and family level characteristics of the child;

$\bar{\mathbf{X}}_{icd} = (N_{cd} - 1)^{-1} \sum_{j \in U_{cd} - i} \mathbf{x}_{jcd}$ is the vector of peer population means in the class;

[4] In a companion paper (Micklewright, Schnepf, and Silva 2012), we investigate the size of the attenuation bias in a particular setting by comparing estimates obtained when peer measures are calculated with the survey sample with those obtained with data on the population from which the sample was drawn. A parallel literature in statistics is concerned with estimates from multilevel models applied to survey data with a hierarchical structure when measures of variables at a higher level are formed by averaging the characteristics of units at a lower level (Woodhouse et al 1996, Kravdal 2006).

$\mathbf{z}_{cd}$ is a vector of teacher and classroom level characteristics; and

$y_{icd}^{t-1}$ is the outcome for the child measured at a previous time point before exposure to the peer environment for which estimated effects are required.

In the model (1) the beta coefficients and the school intercepts $\alpha_d$ are all assumed fixed, but the classroom effects $\eta_{cd}$ and the individual 'idiosyncratic' terms $\varepsilon_{icd}$ are considered random, with mean zero and unspecified but fixed variances $\sigma_\eta^2$ and $\sigma_\varepsilon^2$. The key parameter of interest is $\beta_1$, namely the peer effect.

The model (1) is formulated using population averages for the peers, but only a sample of pupils in each class is available to fit the model. Therefore some of the covariates are 'measured with error', namely the sampling error of estimating the class peer averages from a sample of peers in each class. In the appendix of their paper, NW adopt a largely simplified model given by

$$y_{icd}^t = \beta_1 \bar{W}_{icd} + \varepsilon_{icd} \qquad (2)$$

This is likely to be too simple a model for any application, but nevertheless was the model assumed by NW when developing their simple measurement error adjustment factor in their appendix. NW refer to the model with fixed school effects considered by Ammermueller and Pischke (2006) which is similar to their main model described by (1), but then in a footnote they confirm that their development is based on the simpler 'bivariate regression' model (2).

In the scenario considered by NW, the data available to fit the model (2) are observations made for a sample of peers in each class, and therefore $\bar{W}_{icd}$ is not known. Instead it is replaced by the sample average of peers $\bar{w}_{icd} = (n_{cd} - 1)^{-1} \sum_{j \in s_{cd} - i} w_{jcd}$ , where $s_{cd}$ is the sample of children and $n_{cd}$ is the number of children in the sample of the class.[5] Despite the fact that the model (2) refers to three levels (pupils, classes and schools), the

[5] A quick comment about notation is due here. NW label their population and sample peer effect variables as $w_{cd}$ and $\overset{*}{w}_{cd}$ , respectively, dropping the child's index. However, these variables vary with the children in each class, and therefore the child's index *i* should not be removed.

model structure only recognises two of these levels, since peers are defined within each class and no explicit structure is imposed to capture the school effects.

Let $\hat{\beta}_{1,OLS}$ denote the OLS estimator of $\beta_1$ under (2), but calculated from the available sample data, namely:

$$\hat{\beta}_{1,OLS} = \frac{\sum_{icd}(y_{icd} - \bar{y})(\bar{w}_{icd} - \bar{w})}{\sum_{icd}(\bar{w}_{icd} - \bar{w})^2} = \frac{\text{cov}(y_{icd}, \bar{w}_{icd})}{\text{var}(\bar{w}_{icd})} \tag{3}$$

where $\bar{y}$ and $\bar{w}$ are the overall sample averages of the response and the peer effects variables, respectively.

NW then state that the OLS estimate of $\beta_1$ converges to:

$$\begin{aligned} \text{plim}\hat{\beta}_{1,OLS} &= \frac{COV(y_{icd}, \bar{w}_{icd})}{VAR(\bar{w}_{icd})} \\ &= \frac{COV(\beta_1 \bar{W}_{icd} + \varepsilon_{icd}, \bar{w}_{icd})}{VAR(\bar{w}_{icd})} \\ &= \beta_1 \frac{COV(\bar{W}_{icd}, \bar{w}_{icd})}{VAR(\bar{w}_{icd})} + \frac{COV(\varepsilon_{icd}, \bar{w}_{icd})}{VAR(\bar{w}_{icd})} \end{aligned} \tag{4}$$

Now let’s explore (4). NW drop the second term on the right hand side of the last expression. Indeed a standard (though unspecified) assumption of the model (2) is that $COV(\bar{W}_{icd}, \varepsilon_{icd}) = 0$, namely that the model errors ($\varepsilon_{icd}$) are uncorrelated with the true peer covariate ($\bar{W}_{icd}$).

However the term dropped from (4) is the covariance between the model errors ($\varepsilon_{icd}$) and the sample average peer effects ($\bar{w}_{icd}$), i.e. the peer covariate measured with error. This will be zero only under the additional assumption that the measurement errors $(\bar{w}_{icd} - \bar{W}_{icd})$ are uncorrelated with the true measurements $(\bar{W}_{icd})$, namely $COV(\bar{w}_{icd} - \bar{W}_{icd}, \bar{W}_{icd}) = 0$, and with the model errors ($\varepsilon_{icd}$), namely $COV(\bar{w}_{icd} - \bar{W}_{icd}, \varepsilon_{icd}) = 0$, assumptions generally required by the classical measurement error model specification – see for example Fuller (1987, eq. 1.1.3). However NW make none of these assumptions explicitly. In fact in our own empirical work using data from the Programme for International Student Assessment (PISA) , we estimated the correlation $CORR(\bar{w}_{icd} - \bar{W}_{icd}, \bar{W}_{icd})$ as -0.18 (significantly different from zero using a 0.1 percent significance level) – see (Micklewright, Schnepf and Silva 2012).

The classical measurement error model would also require the assumption that the vectors $(\bar{W}_{icd}, \bar{w}_{icd} - \bar{W}_{icd}, \varepsilon_{icd})$ are independent across different individuals. Now the assumption that both the population and sample averages of peers are uncorrelated across pupils from different schools is easily justified, but the same does not apply if pupils belong to the same school.

NW then derive the variance which appears in the denominator of the terms on the right hand side of (4) as:

$$\begin{aligned} VAR(\bar{w}_{icd}) &= VAR\left[(n_{cd}-1)^{-1}\sum\nolimits_{j\in s_{cd}-i} w_{jcd}\right] \\ &= (n_{cd}-1)^{-2} VAR\left(\sum\nolimits_{j\in s_{cd}-i} w_{jcd}\right) \\ &= (n_{cd}-1)^{-2}\sum\nolimits_{j\in s_{cd}-i} VAR(w_{jcd}) \\ &= (n_{cd}-1)^{-1} VAR(w_{jcd}) \end{aligned} \qquad (5)$$

Note that to obtain the third line above it is essential to assume that sampling of children in each class is with replacement and equal probabilities, so that the independence of the terms in the sum warrants using the property that the variance of a sum is equal to the sum of the variances of the terms. This assumption is not mentioned by NW, and in practice sampling of pupils within classes is never done with replacement. For the data considered by NW it is not even a satisfactory approximation, since pupils are sampled without replacement and the average sampling fraction would not be far from 8.55 / 20.51 = 41.7 percent – see page 13 of NW.

In addition, the plim result provided in (4) refers to model variances and covariances of the variables involved across the whole population. Expression (5) starts by calculating a variance due to sampling of pupils within a specific class and school, and therefore, it does not reflect a property of the variable across the whole population. Notice that the result depends on the specific sample size within each class ($n_{cd}$) and these are not assumed to be constant.

NW then derive the covariance which appears in the numerator of the first term on the right hand side of (4) as:

$$
\begin{aligned}
COV(\bar{W}_{icd}, \bar{w}_{icd}) &= COV\left( \frac{\sum_{j\in U_{cd}-i} w_{jcd}}{N_{cd}-1}, \frac{\sum_{k\in s_{cd}-i} w_{kcd}}{n_{cd}-1} \right) \\
&= \frac{1}{N_{cd}-1}\frac{1}{n_{cd}-1}\sum_{j\in U_{cd}-i}\sum_{k\in s_{cd}-i} COV(w_{jcd}, w_{kcd}) \\
&= \frac{1}{N_{cd}-1}\frac{1}{n_{cd}-1}(n_{cd}-1)\, VAR(w_{jcd}) \\
&= \frac{1}{N_{cd}-1} VAR(w_{jcd})
\end{aligned} \tag{6}
$$

To get the second and third lines of (6) the assumption of independent observations for the covariate *w* within a class is required, but not stated. Once again the result obtained depends on the specific population size within a class ($N_{cd}$).

Now recall that to drop the second term in the last line of (4) the following assumption was required:

$$
COV(\bar{w}_{icd} - \bar{W}_{icd}, \bar{W}_{icd}) = 0 \tag{7}
$$

If this is the case, then it implies that

$$
\begin{aligned}
&COV(\bar{w}_{icd} - \bar{W}_{icd}, \bar{W}_{icd}) = COV(\bar{w}_{icd}, \bar{W}_{icd}) - VAR(\bar{W}_{icd}) = 0 \\
&\Leftrightarrow COV(\bar{w}_{icd}, \bar{W}_{icd}) = VAR(\bar{W}_{icd})
\end{aligned} \tag{8}
$$

Therefore we have a contradiction between (6) and (8). This means that either we adopt the additional hypothesis required to drop the second term on the last line of (4), in which case the derivation of (6) leads to contradictory results, or if we do not adopt this additional hypothesis, we cannot drop this term from (4). In either case, the consequence is that the simple adjustment factor derived does not follow.

From expressions (5) and (6) NW obtain:

$$
\text{plim}\hat{\beta}_{1,OLS} = \frac{n_{cd}-1}{N_{cd}-1}\beta_1
$$

which looks odd, because the probability limit for the OLS estimator for $\beta_1$ depends on the size of class and the sample size for a specified class *c* in school *d*. This result is not reasonable, since the size of the atennuation bias should not depend on how large the sampling fraction is within a sincle class or school. NW hint at the problem when they move from this expression to one where the OLS estimator is adjusted by the average values of the class population and sample sizes, namely:

$$\hat{\beta}_{1,adj} = \frac{\bar{N}_{cd} - 1}{\bar{n}_{cd} - 1} \hat{\beta}_{1,OLS}$$

NW saw that the adjustment should not depend on a single class, but did not recognize that their probability limit suffered from this problem.

Hence our conclusion is that the simple adjustment proposed by Neidell and Waldfogel (2008) for the measurement error induced by sampling of peers is not properly justified and should not be used ‘as is’.

## References

Ammermueller A and Pischke J-S (2006) ‘Peer effects in European Primary Schools: Evidence from PIRLS’ ZEW discussion paper 06-027.

Ammermueller A and Pischke J-S (2009) ‘Peer effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study’ *Journal of Labor Economics* 27(3): 315-48.

Fuller, W. A. (1987) *Measurement Error Models*. Hoboken, NJ: Wiley.

Kravdal O (2006) ‘A simulation-based assessment of the bias produced when using averages from small DHS clusters as contextual variables in multilevel models’ *Demographic Research* 15(1): 1-20.

Micklewright, J, Schnepf, S V, and Silva P N (2012) ‘Peer effects and measurement error: the impact of sampling variation in school survey data (evidence from PISA)’ *Economics of Education Review* 31 (6): 1136–1142.

Neidell M and Waldfogel J (2008) ‘Cognitive and non-cognitive peer effects in early education’ NBER working paper 14277.

Neidell M and Waldfogel J (2010) ‘Cognitive and non-cognitive peer effects in early education’ *Review of Economics and Statistics*, 92(3): 562-576.

Woodhouse G, Yang M, Goldstein H, and Rasbash J (1996) ‘Adjusting for measurement error in multilevel analysis’, *Journal of the Royal Statistical Society,* Series A, 159 (2): 201-212.