Sibling Models, Categorical Outcomes, and the Intra-Class Correlation

Richard Breen [1,2] and John Ermisch [1,3]

1.Nuffield College, University of Oxford

2.Department of Sociology, University of Oxford

3.Leverhulme Centre for Demographic Science, University of Oxford

Thursday, 1 October 2020

We would like to thank the reviewers for their helpful comments on earlier drafts and Colin Mills for bringing to our attention the sibling data in the Oxford Mobility Study.

Breen acknowledges support from ESRC grant ES/P009301/1 , "Understanding Inequalities".

Sibling Models, Categorical Outcomes, and the Intra-Class Correlation

Abstract: In sibling models with categorical outcomes the question arises of how best to calculate the intraclass correlation, ICC. We show that, for this purpose, the random effects linear probability model is preferable to a random effects non-linear probability model, such as a logit or probit. This is because, for a binary outcome, the ICC derived from a random effects linear probability model is a non-parametric estimate of the ICC, equivalent to a statistic called Cohen's κ. Furthermore, because κ can be calculated when the outcome has more than two categories, we can use the random effects linear probability model to compute a single ICC in cases with more than two outcome categories. Lastly, ICCs are often compared between groups to show the degree to which sibling differences vary between groups: we show that when the outcome is categorical these comparisons are invalid. We suggest alternative measures for this purpose.

Sibling Models, Categorical Outcomes, and the Intra-Class Correlation

In recent years there has been a renewed interest among sociologists in sibling models (early examples include Hauser and Featherman 1976, Jencks 1979, and deGraaf 1986); this interest has extended to other social scientists too (Bjorklund et al 2002 among others). Recent applications in the field of educational sociology, for example, include studies of sibling correlations in cognitive ability (Duncan, Boisjoly, and Harris, 2001; Björklund and Jännti 2012), years of education (Sieben and de Graaf 2001; Marks and Mooi-Reci 2016), and school attainment (Rasbash et al., 2010; Nicoletti and Rabe 2018). The major value of sibling models is that they allow us to calculate the overall variation in an outcome that is explained by characteristics shared by siblings, including genes, parental resources, and neighbourhood features (Breen and Jonsson 2005: 232-3), and so they are used to study the overall impact of family background on siblings' outcomes, rather than the effect of one particular aspect, such as parental income or parental social class.

In studying sibling data (with twin data being a particular case) the intra-class correlation, the ICC, is widely used. When the outcome of interest is continuous and we assume that the variance within sibling pairs is independent of the variance between them, the ICC, calculated as the ratio of the between-family variance to the overall variance, is equal to the sibling correlation in the outcome. The ICC is a convenient measure of the proportion of variation in the outcome that lies between sibling pairs; the larger this is, the greater the strength of common (family) influences. ICCs for continuous outcomes are also often compared between groups, such as nations (Grätz et al. 2019), parental background (Conley and Glauber, 2008), or both (Grätz et al. 2019). Some theories argue that educational attainment should be more similar among siblings from more advantaged families because they can afford to invest in the weaker-performing sibling to try to equalise outcomes. This is sometimes called a compensatory strategy (Conley, 2008). Greater variation between siblings from less advantaged origins might also arise if, because of limited resources, families invest in the child which promises the greatest returns (Behrman, Pollak, and Taubman 1982).

However, when the outcome of interest is binary (or, in general, categorical), rather than continuous, matters are more complicated in respect of both the calculation of the ICC and making comparisons of ICCs. In this note we have three goals. First, we show that the random effects linear probability model fitted to a binary outcome is, in fact, a non-parametric estimator of the ICC, equivalent to a statistic called Cohen's κ (Cohen 1960). This makes the random effects linear probability model preferable to a random effects non-linear probability model (typically a multilevel logit or probit model) for calculating the ICC. Secondly, we show that, because κ can be calculated when there are more than two outcomes, we can generate a single ICC in such cases using random effects linear probability models. Finally, we show that the common practice of comparing ICCs for different groups is invalid when the outcome is categorical. We suggest the use of the within-family variance instead, or, equivalently, a measure we develop later and call $1 - q$ (equal to the proportion of families in which the siblings have different outcomes).

Although sibling models are the focus of this paper, the points we make apply more widely to situations in which we want to compute ICCs between groups of any kind (not only sibling pairs) in respect of outcomes of any kind that are categorical rather than continuous. For sociologists, social classes spring to mind: given sets of unordered social classes as both origin and destination we might want to compute the ICC of origin classes with respect to destination classes. This would tell us how much of the variation in class attainment lay between people from different class origins and how much lay among people from the same origin class. In political sociology we might ask how much variation in people's party identification lies between different regions of the country and how much within them. This might give some insight into the social bases of party support. An obvious extension in the field of education would be to estimate the variation between schools, compared with the variation within them, in choice of post-secondary destinations, where the alternatives are treated as categories.

In our exposition, we focus on sibling data. We assume that we have data consisting of sibling pairs, nested within families. We begin by considering a binary outcome: $Y_{ij} = 1 \; or \; 0$ where $i = 1, \ldots, N$ denotes families and $j = 1, \ldots, n_i$ denotes individual siblings within a family. Later we deal with outcomes having more than two categories. In our exposition $n_i = n = 2$ for all families. In multilevel terminology, siblings are level-1 observations, families are level-2.

The Intra-class Correlation

The ICC is defined as the ratio of the variance in the outcome between families to the total variance in the outcome. The total variance is equal to the sum of the between-family variance and the within-family (between sibling) variance. In a random effects linear model the within-variance is estimated by

$$\sigma_e^2 = \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 / N(n-1) \qquad (1)$$

where $\bar{y}_i$ is the mean within family *i*.

The total variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2}{Nn} \qquad (2)$$

Here $\bar{y}$ is the overall mean. The ICC can be written $1 - \frac{\sigma_e^2}{\sigma^2}$. However, random effects non-linear probability models do not estimate σ_e^2: instead the level-1 variance is fixed at the variance of the standard version of whatever distribution has been chosen for the latent variable. In the random effects logit model the level-1 variance is fixed at $\pi^2/3$ and this remains unchanged, regardless of whether variables are added to the model that might account for some of the level-1 variation (see, for example, Guo and Zhao 2000: 451; Snijders and Bosker 2004, chapter 14).

To show, for the case where the outcome is binary, that equations 1 and 2 yield a non-parametric estimator of the ICC, first define q as the proportion of all families in which the siblings' outcomes are the same. Now consider applying equation 1 to calculate the within-family variance. For any family in which both siblings have the same outcome, $y_{ij} - \bar{y}_i = 0$ but if the siblings have different outcomes, $y_{ij} - \bar{y}_i = \pm 1/2$ and so $\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 = 1/2$. The within variance is thus equal to one half times the proportion of families in which the siblings differ in their outcomes, that is $\frac{1-q}{2}$.

Let p denote the proportion of observations that have $y_{ij} = 1$. Then $\bar{y} = p$. Because Y is binary, the total variance is $p(1-p)$. This is equal to the total variance given in equation 2 because the proportion of individuals with *Y = 1* is $\bar{y}$ and they contribute $(1-\bar{y})^2$ to the total variance. The proportion of individuals with *Y = 0* is $1-\bar{y}$ and they contribute $\bar{y}^2$ to the total variance. This gives the total variance = $\bar{y}(1-\bar{y})^2 + (1-\bar{y})\bar{y}^2$ and after some algebra this reduces to $\bar{y}(1-\bar{y})$ or $p(1-p)$.

We find, therefore, that the ICC equals

$$1 - \frac{\sigma_e^2}{\sigma^2} = 1 - \frac{1-q}{2p(1-p)} \qquad (3)$$

Since q and p are calculated directly from the data, this demonstrates that the random effects linear probability model is a non-parametric estimator of the ICC.

We can use the delta method (Oehlert 1992) to calculate the standard error of the ICC. Because p and q are independent the formula is

$$s.e.(ICC) = \sqrt{\left[\frac{\partial ICC}{\partial p}\right]^2 \sigma_p^2 + \left[\frac{\partial ICC}{\partial q}\right]^2 \sigma_q^2}$$

With $\sigma_p^2 = p(1-p)/N, \sigma_q^2 = q(1-q)/N$ and $\frac{\partial ICC}{\partial p} = \frac{2(1-ICC)^2(1-2p)}{1-q}$ and $\frac{\partial ICC}{\partial q} = \frac{1-ICC}{1-q}$.

Cohen's κ

The expression we derived for the ICC is also the formula for Cohen's (1960) κ in the case where $n = 2$. κ is a measure of agreement in the outcome within level-2 units, corrected for agreement that would be expected by chance. κ was developed to address questions of inter-rater reliability, measuring the degree to which independent judges categorized subjects in the same or a different way. Compared with sibling studies, the subjects are equivalent to families (level-2 units) and the judges to siblings (level-1 units). Fleiss (1971) extended κ to deal with cases in which there were several sets of judges, but each of the same number: for example, several pairs of judges rather than the single pair considered by Cohen. Fleiss and Cuzick (1979) extended this to the case in which the number of judges of each subject can vary; this is equivalent to multi-level data with different numbers of siblings nested within families.

A general expression for κ can be derived as follows. Because we now allow for more than two categorical outcomes, we index outcome categories by $K = 1, \dots, K$. Define h_{ik} as the number of siblings in the i^{th} family with $y_{ij} = k$. The proportion of pairs of siblings whose outcome are the same, out of all the possible pairs, in the i^{th} family is

$$P_i = \frac{1}{n(n-1)} \sum_{k=1}^{K} h_{ik}(h_{ik} - 1) \qquad (4)$$

The mean proportion of agreeing pairs is then

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i \qquad (5)$$

But some agreement in the outcomes of the siblings would be expected by chance, depending on the distribution of *Y*. This is equal to

$$\bar{P}_e = \sum_{k=1}^{K} p_k^2 \qquad (6)$$

where p_k is the overall proportion in the k^{th} outcome category.

Then,

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \qquad (7)$$

"The quantity $1 - \bar{P}_e$ measures the degree of agreement attainable over and above what would be predicted by chance. The degree of agreement actually attained in excess of chance is $\bar{P} - \bar{P}_e$, so … [κ is] … a normalized measure of overall agreement, corrected for the amount expected by chance" (Fleiss 1971: 379, parentheses added).

In the case in which *n = 2* and $k = 2$, $\bar{P} = q$, $\bar{P}_e = 1 - 2p(1-p)$ and so $\kappa = 1 - \frac{1-q}{2p(1-p)}$, exactly equal to the ICC (see equation (3)). In fact, we can immediately see the link with the ICC when we consider that *1-q* can also be interpreted as the probability that a randomly chosen person has a sibling with a different value of *Y* while $2p(1-p)$ is the probability of choosing any pair of people whose outcomes differ. Landis and Koch (1977) derived an ICC by applying a one-way analysis of variance to binary data, and Fleiss and Cuzick (1979: 538-9) show that this is identical to κ, except for a correction factor of $\frac{N-1}{N}$ in the computation of the between mean square of the ANOVA approach. But, clearly, if N (the number of level-2 units) is even moderately large, this correction factor approaches 1.[1]

Example: Sibling Data

We illustrate the statistics with the following example. We match adult siblings from the British Household Panel Survey (BHPS) data who are aged 22-40 at their last observation in the panel. Using only families in which we identify exactly two siblings, there are 226 sibling pairs. The outcome variable is whether the person has a University degree or not, with $E(Y) = 0.34$ and $var(Y)$ = 0.224 and q = 0.721 (s.e. = 0.030). κ is estimated as 0.3775 (s.e. = 0.066), the ICC from a random effects linear probability model is 0.3794 (the standard error, computed using the delta method shown earlier, is 0.070). The agreement between κ and the random effects linear probability model estimate of the ICC is very close: the small difference between them disappears once the degrees of freedom correction is made. By contrast, the ICC from a random effects logit model is 0.54. Unlike the ICC from the random effects linear probability model or κ, the ICC from the random effects logit model refers to a hypothetical latent, logistically distributed latent variable.

More than two outcomes

Fleiss (1971) derived κ for the case in which there are more than two outcomes and then showed that the overall κ is equal to a variance-weighted average of the separate κs for each response

[1] Despite the equivalence between the ICC from the random effects LPM and κ, we believe that most sociologists will find the former method easier and more intuitive; certainly it is more familiar. In addition, the random effects LPM lends itself easily to the calculation of conditional ICCs, controlling for predictor variables.

category. Using the same data, Landis and Koch (1977; see also Fleiss and Cuzick 1979) computed an overall ICC as a variance weighted average of the ICCs for each outcome.

Because κ for more than two outcomes is a weighted sum of κs for binary outcomes, random effects linear probability models can be used to estimate an overall κ which can be thought of as an overall ICC. Given outcomes, $k = 1,\ldots,K$, the random effects model is fitted to each of the K pairwise contrasts ($Y_k = 1 \; vs. Y_k = 0 \;\; \forall \, k$), the ICC is calculated, and the overall ICC or κ is given by

$$ICC = \kappa = \sum_{k=1}^{K} ICC_k \frac{var(Y_k)}{\sum_k var(Y_k)} \qquad (8)$$

Thus the random effects linear probability model can be used to generate a non-parametric estimate of the ICC when there are more than two categorical outcomes.

The data in Table 1 were analysed by Fleiss (1971, Table 1) and reanalysed by Landis and Koch (1977, Table 2). Subjects are level-2 units and they were diagnosed by six raters into one of five categories. The table shows the number of ratings in each category. In this case, Y has five categories and there are 6 level-1 observations in each level-2 unit. Fleiss (1971) reports an estimate of κ of 0.43 (se = 0.028) and Landis and Koch (1977) an estimate of the ICC of 0.44 (se = 0.054).

[Table 1 here]

These approaches are exactly equivalent to running a set of binary random effects linear probability models in which each outcome category is compared with all the others, computing the ICC for each, and then forming a variance-weighted ICC. Table 2 shows our calculations using the random effects LPM and κ (these can be compared with the estimates reported in Landis and

Koch 1977, Table 3) applied to the data in Table 1. We computed the standard errors of the overall ICC and κ using the delta method. The two sets of estimates agree very closely.

[Table 2 here]

Because κ measures agreement and disagreement, in calculating the overall κ as the weighted sum of binary outcome κs we are treating all forms of disagreement in the same way. For example, if our outcomes were social classes, the case in which one sibling was in the highest class and the other was in the second highest class would be treated in exactly the same way as if the second sibling was in the lowest class. Both would be examples of disagreement. But we might instead want to treat the disagreement as greater in the second case than in the first. This can be accomplished by weighting, yielding a weighted κ or ICC. In the case of social classes, we might define weights as a function of the "distance" between the classes. But there may be other weighting schemes and, unless one of them is compelling, perhaps the best course of action is to compute a range of κs using different weightings to generate some bounds on the estimate.

As an illustration, we consider the data in Table 3, taken from the Oxford Mobility Study (University of Oxford 1977). This was a survey of men and the respondents were asked to provide information about a randomly chosen brother. Their responses have been coded into the seven Goldthorpe classes, ranging from I Higher Professional and Managerial to VII Semi-skilled and Unskilled Manual Workers (see Goldthorpe 1980: 40-43 for the full schema). Table 4 shows estimates of the ICC for the unweighted case and for four different weighting schemes. The first uses the row and column scores from a homogenous (equal row and column scores) RC2 log-multiplicative row and column effects model (Goodman 1979) with the main diagonal of the table fitted exactly.[2] These scores maximize the association between respondents' and brothers' class position: thus, this is the scoring that comes closest to maintaining the pattern of association

[2] This model fits the data shown in Table 4 with deviance = 36.63, df = 23, p = .036.

in the data. In the case of DW1 we rank the classes from I to VII and compute a distance weight as $1 - |i - j|/(k - 1)$, where *i* and *j* index the rows and columns of the social class rank of the two brothers and k is the maximum number of social classes. DW2 uses the same rankings but computes the distance by $1 - \left\{\frac{|i-j|}{k-1}\right\}^2$; these weights make the categories more alike than DW1. Lastly, the column headed "AW" ("adjacent weights") gives adjacent social classes a weight of 0.9 and all others a weight of zero.

[Tables 3 & 4 here]

The unweighted ICC (0.15) suggests that England and Wales are rather open societies inasmuch as the great majority of the variation in class position is between siblings, rather than between families. The other estimates return an ICC that is larger, though, they too would suggest that by far the larger share of the variance lies between siblings rather than between families. RC2 and DW1 give very similar estimates, being about twice the unweighted ICC. They are similar because the row and column scores estimated under the RC2 model are very close to the rank ordering of classes. For DW1, for example, when the brothers' social classes differ by 2 classes (i.e. there is an intervening social class) the distance is 0.6667, meaning that the brothers' social classes are in two-thirds agreement. The ICC using DW2 is larger: this is not surprising because making the classes more alike (which is what the DW2 weighting does) must increase the between-class variation. The ICC from AW is smaller because all pairs in non-adjacent classes are considered to be in disagreement. And, of course, the unweighted ICC is the smallest because all cases where the brothers are in different classes are considered as disagreements.

Comparing the ICC across groups

Some studies using siblings or twins compare ICCs calculated for different groups (for example, Karhula 2015, Senghor and Wolff 2017). In studies focusing on educational outcomes, the ICCs of children from different social origin groups are sometimes compared: "The analysis of multilevel models restricted to specific social origin groups tests whether sibling similarity varies by family socioeconomic background" (Grätz 2018: 255). More generally, many studies compare group-specific ICCs to establish whether "certain family structures exhibit more 'clustering' of sibling behavioral and developmental outcomes than others" (Conley et al 2007: 1090). In twin studies, for example, the ICCs of same-sex or mixed-sex twin pairs might be compared or the ICCs of MZ and DZ twins (Conley et al. 2013; Duncan, Boisjoly, and Harris 2001).

Comparing ICCs in this way when they are computed from categorical outcomes is problematic. For binary outcomes, equation 3 shows that, when the total variance is large, all else equal, the ICC will be big and when it is small the ICC will be small. But because the variance of a binary variable depends on its mean, the ICCs of different groups can differ because of differences in the proportion of observations in the group with $Y = 1$ or because the within-family variation differs, or both. Consider the case of two groups, both with the same within-group variation, $(1 - q)/2$. If, however, 20% of group A have $Y = 1$ compared with 50% in group B, the ICC for group A will be $1 - \frac{1-q}{0.32}$ while for group B it will be $1 - \frac{1-q}{0.50}$. The ICC will be larger in group B. But if our concern is with group differences in sibling similarity the ICC would obviously be misleading because, in this case, it differs between the two groups only because of group differences in the average outcome, not group differences in average within-family variability. This problem applies irrespective of whether a random effects linear or non-linear probability model is used but in the latter case the within-family variance is, in any case, fixed and so group differences in the ICC can only reflect differences in the between-family variance. In our view it would be better to focus on the within-group variance itself or on *1-q*; both are more informative than the ICC in telling us if different groups display more or less heterogeneity of outcomes among siblings within the same families. [3]

[3] The problem arises because, for binary variables, the mean and variance are not separately identified. But, as pointed out to us by a reviewer, even in the case of a continuous outcome the ICC might still obscure the real quantity of interest because it is a ratio measure. If our concern is to compare the variation within

When there are multiple outcome categories, similar arguments apply. As we showed above, the ICC for the case in which there are more than two outcome categories is equal to the general formula for $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. κ depends on both the mean proportion of agreeing pairs ($\bar{P}$) and the average agreement across categories ($\bar{P}_e$) in each group. Two groups with the same $\bar{P}$ could have different values of κ because $\bar{P}_e$ differed between the groups. But, when there are two siblings in each family, $\bar{P}$ is equal to q regardless of how many outcome categories there are. This follows because, with two siblings, h_{ik} (the number of siblings in outcome k) is always 0, 1 or 2 and thus P_i in equation (4) is 1 when both siblings are in the same category, 0 otherwise. Then, $\bar{P}$ is simply the proportion of families in which $P_i = 1$; in other words, the proportion of families in which both siblings' outcomes are the same. But that is our definition of q. If we cross-tabulate the outcomes of the two siblings (as in Table 3), q (equally, $\bar{P}$) is the proportion of cases on the main diagonal and $\bar{P}_e$ is the proportion of cases that would be expected to lie on the main diagonal if the siblings' outcomes were independent. So, to rephrase: two groups with the same q could have different ICCs because the marginal distributions of the outcomes differed between the groups.

To illustrate this we use the samples of brothers from the Oxford Mobility Study to make comparisons of q and the unweighted ICC according to the social class of the father. In other words, our data consists of a table of the same form as Table 3 but now for each of the seven father's social classes. As Table 5 shows, the ranking of similarity between brothers by father's social class differs, depending on which of these two measures is used. Brothers are most heterogeneous in father's social class III on the basis of q (they have the largest $1 - q$) but, on the basis of their ICCs, brothers are most heterogenous in father's social class VII (they have the smallest ICC). Furthermore, father's social classes V, VI and VII are all more heterogenous than III according to their ICCs, but the opposite is true according to q. Of course, these estimates are

---

families from different groups we might want to report the average within-family variance in each group as well as, or instead of, the group ICCs.

not very precise, so not too much should be made of this difference. The confidence intervals for the ICCs do, however, indicate that there is more similarity between brothers in father's classes I and II then there is in the lower social classes. The main point, however, is that q is, in our view, a better measure of within-family similarity than the ICC but researchers would be advised to report both, just as they should show both the ICC and the within-family variance when the outcome variable is continuous.

Conclusion

We have shown that the random effects linear probability model yields a non-parametric estimate of the ICC: in our view this makes it preferable to the ICC computed from a random effects logit (or other non-linear probability) model which, in any case, rests on untestable assumptions about the variance within sibling pairs. Second, we showed that the random effects linear probability model can be used to estimate an ICC in cases where there are more than two categorical outcomes, though this will normally require the use of weights, the choice of which will have to be justified. Finally, we have also shown that comparing ICCs for different groups is potentially fraught with problems. If our interest really lies in uncovering variations in within-family outcomes across different groups it would be more straightforward to use a direct measure of within-family variation.

References

Behrman, J. R, Pollak, R. A. and Taubman, P. 1982. Parental preferences and provision for progeny. Journal of Political Economy, 90: 52-73.

Björklund, Anders and Jäntti, Markus. 2012. How important is family background for labor-economic outcomes? Labour Economics, 19: 465-474.

Björklund, Anders, Tor Eriksson, Markus Jantti, Oddbjorn Raaum, and Eva Osterbacka. 2002. Brother Correlations in Earnings in Denmark, Finland, Norway, and Sweden Compared to the United States. Journal of Population Economics 15: 757-72.

Breen, Richard and Jan O, Jonsson. 2005. Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility. Annual Review of Sociology 31: 223-244.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20,1: 37-46.

Conley, Dalton. 2008. Bringing Sibling Differences In: Enlarging Our Understanding of the Transmission of Advantage in Families, in: A. Lareau and D. Conley Social Class: How Does It Work? (New York, Russell Sage Foundation)

Conley, Dalton, & Glauber, Rebecca. 2008. All in the family?: Family composition, resources, and sibling similarity in socioeconomic status. Research in Social Stratification and Mobility, 26: 297-306.

Conley, Dalton, Kathryn M. Pfeiffer and Melissa Velez. 2007. Explaining sibling differences in achievement and 17eighbour17l outcomes: The importance of within- and between-family factors. Social Science Research 36: 1087–1104.

Conley, Dalton, Emily Rauscher, Christopher Dawes, Patrik K. E. Magnusson and Mark L. Sigal. 2013. Heritability and the Equal Environments Assumption: Evidence from Multiple Samples of Misclassified Twins. Behavior Genetics 43,5:413-26.

deGraaf, Paul M. 1986. The Impact of Financial and Cultural Resources on Educational Attainment in the Netherlands. Sociology of Education 59: 237-246.

Duncan, G. J, Boisjoly, J, & Harris, K. M. 2001. Sibling, peer, 17eighbour, and schoolmate correlations as indicators of the importance of context for adolescent development. Demography, 38: 437-447.

Fleiss, Joseph L. 1971. Measuring Nominal Scale Agreement among Many Raters. Psychological Bulletin 76,5: 378-82.

Fleiss, Joseph L. 1975. Measuring Agreement between Two Judges on the Presence of Absence of a Trait. Biometrics 31, 3: 651-59.

Fleiss, Joseph L. and Jack Cuzick. 1979. The Reliability of Dichotomous Judgments: Unequal Numbers of Judges per Subject. Applied Psychological Measurement, 3, 4: 537-42.

Goldthorpe, John H. 1980, Social Mobility and Class Structure in Modern Britain. Oxford: Clarendon Press.

Guo, Guang and Hongxin Zhao. 2000. Multilevel Modeling for Binary Data. Annual Review of Sociology 26: 441-64.

Grätz, Michael. 2018. Competition in the Family: Inequality between Siblings and the Intergenerational Transmission of Educational Advantage. Sociological Science 5: 246-69.

Grätz, M., Barclay, K., Wiborg, Ø.N., Lyngstad, T.H., Karhula, A., Erola, J., Präg, P., Laidley, T. and Conley, D. 2019. Universal family background effects on education across and within societies, Stockholm Research Reports in Demography, No. 2019:13.

Goodman, Leo A. 1979. Simple Models for the Analysis of Association in Cross-Classifications having Ordered Categories. Journal of the American Statistical Association 74: 537-52

Hauser, Robert M. and David L. Featherman. 1976. Equality of schooling: trends and prospects. Sociology of Education 49: 99-120.

Jencks, Christopher et al. 1979. Who Gets Ahead? The Determinants of Economic Success in America. New York: Basic Books.

Karhula, Aleksi. 2015. Comparing Overall Effects of Family Background on Homeownership During Early Life Course. Housing Studies, 30: 1281-1298.

Landis, J. Richard and Gary G. Koch. 1977. A One-Way Components of Variance Model for Categorical Data. Biometrics 33, 4: 671-9.

Marks, G. N, and Mooi-Reci, I. 2016. The declining influence of family background on educational attainment in Australia: The role of measured and unmeasured influences. Social Science Research, 55: 171-185.

Nicoletti, C, and Rabe, B. 2018. Sibling spillover effects in school achievement. Journal of Applied Econometrics, 34: 482-501.

Oehlert, Gary W. 1992. A Note on the Delta Method, The American Statistician, 46: 27-29

Rasbash, J, Leckie, G, Pillinger, R, & Jenkins, J. 2010. Children's educational progress: partitioning family, school and area effects, Journal of the Royal Statistical Society: Series A (Statistics in Society), 173: 657-682.

Senghor, Hady and François-Charles Wolff. 2017. Educational Inequalities between Siblings: Evidence from Six Sub-Saharan African Countries. African Development Review 29: 223–236.

Sieben, I, & De Graaf, P. M. 2001). Testing the modernization hypothesis and the socialist ideology hypothesis: a comparative sibling analysis of educational attainment and occupational status. The British Journal of Sociology, 52: 441-467.

Snijders, Tom A.B. and Roel J. Bosker. 2004. Multilevel Analysis: An introduction to basic and advanced multilevel modelling. Sage, London.

University of Oxford, Oxford Social Mobility Group, Goldthorpe, J. H., Halsey, A. H., Heath, A., Payne, C. (1978). Social Mobility Inquiry, 1972. [data collection]. UK Data Service. SN: 1097, http://doi.org/10.5255/UKDA-SN-1097-1

TABLES

Table 1: Diagnoses on 30 Subjects by Six Raters per Subject (from Fleiss 1971, Table 1)

	Frequency counts of diagnosis				
Subject	Depression	Personality disorder	Schizophrenia	Neurosis	Other
1	0	0	0	6	0
2	0	3	0	0	3
3	0	1	4	0	1
4	0	0	0	0	6
5	0	3	0	3	0
6	2	0	4	0	0
7	0	0	4	0	2
8	2	0	3	1	0
9	2	0	0	4	0
10	0	0	0	0	6
11	1	0	0	5	0
12	1	1	0	4	0
13	0	3	3	0	0
14	1	0	0	5	0
15	0	2	0	3	1
16	0	0	5	0	1
17	3	0	0	1	2
18	5	1	0	0	0
19	0	2	0	4	0
20	1	0	2	0	3
21	0	0	0	0	6
22	0	1	0	5	0
23	0	2	0	1	3
24	2	0	0	4	0
25	1	0	0	4	1
26	0	5	0	1	0
27	4	0	0	0	2
28	0	2	0	4	0
29	1	0	5	0	0
30	0	0	0	0	6

Table 2: ICC and kappa estimated using data from Fleiss 1971, Table 1

	reLPM		κ	
Outcome:	ICC	s.e.(ICC)	κ	s.e.(κ)
Depression	0.254	0.081	0.245	0.047
Personality disorder	0.254	0.081	0.245	0.047
Schizophrenia	0.530	0.082	0.520	0.047
Neurosis	0.481	0.084	0.471	0.047
Other	0.575	0.079	0.566	0.047
Overall	0.440	0.037	0.430	0.022

Table 3: Respondent's class by brother's class, Oxford Mobility Survey

Respondent's class	Brother's class						
	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
I	272	136	67	58	79	77	67
II	117	114	53	57	28	81	91
III	64	65	82	45	67	97	126
IV	74	73	51	179	77	121	132
V	65	81	59	64	153	187	203
VI	66	97	79	111	183	474	380
VII	75	90	107	102	185	373	584

*Table 4: Estimates of the ICC for the data in Table 3 using different weightings**

	Unweighted	RC2	DW1	DW2	AW
ICC	0.1543	0.3014	0.2843	0.3762	0.2506
s.e.	0.0055	0.0090	0.0087	0.0126	0.0009

* Notes:
RC2: a homogenous (equal row and column scores) log-multiplicative row and column effects model.
DW1: classes ranked from I to VII and a distance weight as $1 - |i - j|/(k - 1)$, where *i* and *j* index the rows and columns of the social class rank of the two brothers and *k* is the maximum number of social classes.
DW2: uses the same rankings as DW! but computes the distance by$1 - \left\{\frac{|i-j|}{k-1}\right\}^2$;
AW: gives adjacent social classes a weight of 0.9 and all others a weight of zero.

Table 5: Estimates of the ICC and q for Social Class of Brothers stratified by Father's Social Class, Oxford Mobility Survey data

Father's social class	I	II	III	IV	V	VI	VII
q	0.420	0.340	0.258	0.289	0.264	0.294	0.295
ICC	0.224	0.193	0.130	0.139	0.125	0.113	0.101
s.e.(ICC)	0.026	0.026	0.022	0.015	0.016	0.011	0.012
95% confidence intervals for ICC							
lower	0.172	0.142	0.087	0.110	0.094	0.091	0.077
upper	0.276	0.244	0.173	0.168	0.156	0.135	0.125
N	350	291	372	833	723	1690	1636