MEDIATION ANALYSIS FOR ASSOCIATIONS OF CATEGORICAL VARIABLES: THE ROLE OF EDUCATION IN SOCIAL CLASS MOBILITY IN BRITAIN

BY JOUNI KUHA[1], ERZSÉBET BUKODI[2,*] AND JOHN H. GOLDTHORPE[3,†]

[1]*Department of Statistics, London School of Economics and Political Science, j.kuha@lse.ac.uk*

[2] *Department of Social Policy and Intervention and Nuffield College, University of Oxford, *erzsebet.bukodi@spi.ox.ac.uk*

[3] *Nuffield College, University of Oxford, †john.goldthorpe@nuffield.ox.ac.uk*

We analyse levels and trends of intergenerational social class mobility among three post-war birth cohorts in Britain, and examine how much of the observed mobility or immobility in them could be accounted for by existing differences in educational attainment between people from different class backgrounds. We propose for this purpose a method which quantifies associations between categorical variables when we compare groups which differ only in the distribution of a mediating variable such as education. This is analogous to estimation of indirect effects in causal mediation analysis, but is here developed to define and estimate population associations of variables. We propose estimators for these associations, which depend only on fitted values from models for the mediator and outcome variables, and variance estimators for them. The analysis shows that the part that differences in education play in intergenerational class mobility is by no means so dominant as has been supposed, and that while it varies with gender and with particular mobility transitions, it shows no tendency to change over time.

1. Introduction. In recent years social mobility has become a central political concern in many societies, and notably so in the UK and the US. Under conditions of increasing economic and social inequality, growing attention has centred on rates and trends in mobility between generations. Economists have typically focused on mobility within the income distribution, and sociologists on mobility between social strata defined in various ways. Of late, an increasing amount of research has been carried out on mobility between social classes, defined as collectivities whose members are involved in differing employment relations (see further Goldthorpe 2007, ch. 5). In the UK this understanding of social class is embodied in the categories of the Office for National Statistics Socio-Economic Classification (NS-SEC), which is operationalised through employment status and detailed occupational codes (Rose and Pevalin 2003; Office for National Statistics 2005). In this article we examine social class mobility in the UK over recent decades, using data from the three British birth cohort studies of individuals who were born in 1946, 1958 and 1970.

For this approach, analysis starts from a two-way *mobility table* which cross-classifies individuals by their *class of origin*, their parents' class, and *class of destination*, their own class once they have reached mid-life. What is referred to as *absolute mobility* is defined simply by percentages of change in this table. The total mobility rate is the proportion of individuals whose class destination is different from their class origin, and, to the extent that the classes are ordered, this rate can be decomposed into its 'upward' and 'downward' components. Absolute rates will obviously be influenced by the marginal distributions of origins and destinations, which reflect changes in the overall class structure over time.

Keywords and phrases: Categorical data analysis, finite-population estimation, multinomial logistic models, path analysis

In contrast, *relative mobility* concerns individuals' mobility chances considered net of such structural change. This can be summarized by associations between origins and destinations in the mobility table. They are normally quantified using odds ratios, which compare the conditional probabilities of individuals in one rather than another of two classes of origin being found in one rather than another of two classes of destination. Because odds ratios can vary independently of the marginal distributions of origins and destinations, they are particularly suitable for describing levels of relative mobility. An odds ratio of 1 implies no association or a situation of 'perfect mobility', while ratios further from 1 indicate greater inequality in relative mobility chances, or lower 'fluidity' within the class structure.

Results from a large body of research show that changes in absolute rates of class mobility within countries and differences between countries — both of which can be substantial — are overwhelmingly driven by changes in the class structure, whereas relative rates show a high degree of constancy over time and across countries. Although some variation in these rates occurs, it appears generally slight and unsystematic. In particular, no well-sustained and general equalisation in relative rates over time is apparent (for Britain, see Bukodi and Goldthorpe 2018, for the US, Mitnik, Cumberworth and Grusky 2016, and for cross-national results, Erikson and Goldthorpe 1992, Bukodi, Paskov and Nolan 2020, Bukodi and Paskov 2020, and Breen and Müller 2020).

Moving beyond bivariate analysis of origins and destinations, an important further question is what role is played by education in relative social mobility (see Bukodi and Goldthorpe 2018). It is plausible that education could 'mediate' lack of mobility, if individuals from different class origins have different distributions of educational qualifications and education in turn affects their chances of reaching different classes of destination (see Figure 1, where the arrows indicate the time order in which these characteristics are realised). In political and policy circles, education is widely regarded as being key to increasing social mobility. With educational expansion and reform, it is believed, an 'education-based meritocracy' can be brought into being in which the association between individuals' social origins and educational attainment is reduced, by being made more dependent on 'merit' than on, say, mere accidents of birth, so that the overall association between origins and destinations will also be weakened. However, the degree of constancy in rates of relative mobility that has been observed across very different educational settings must call into doubt whether education can in this way serve as the essential means of breaking the link between inequality of condition and inequality of opportunity.

To put the matter otherwise, what the belief in education as the key to mobility presumes is that the overall association between origins and destinations very largely results from the 'indirect' associations existing between origins and education and then between education and destinations (the latter being taken as constant), rather than the 'direct' association between origins and destinations that is created by differences in families' economic, cultural and social resources operating in other ways than through education. But how far can observed (im)mobility actually be accounted for in terms of class differences in educational attainment, and is any change evident in this regard? These are the sociological research questions that we want to address in this article, for British society over the period which is spanned by the three birth cohorts. In order to answer these questions, we also need to answer the methodological research question of how to define and quantify the part that education does in fact play in social mobility, especially where mobility is treated not in terms of a continuous variable such as income but in terms of social class categories.

In methodological terms, this is a problem in *mediation analysis* or *path analysis*, where we are interested in how the variable M in Figure 1 may mediate the relationship between X and Y. The earliest methods for it were based on combinations of regression models for M and Y. When these are linear models for continuous variables, methods of path analysis go

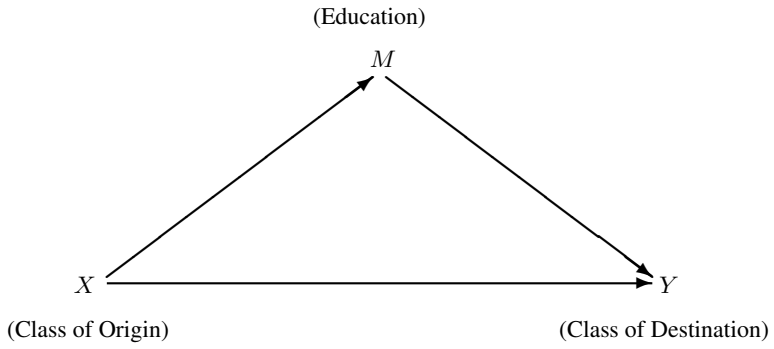


FIG 1. *The setup of basic mediation analysis of three variables. The names of variables in parentheses refer to the social mobility example considered in the article.*

back to Wright (1921), other important early contributors include Tukey (1954) and Blalock (1964), and an influential more recent article is Baron and Kenny (1986); see Wolfle (2003) and Denis and Legerski (2006) for historical reviews, and Bollen (1989) for an overview of this kind of path analysis. In this simple setting, direct and indirect associations (or 'effects') can often be expressed and estimated very simply in terms of the coefficients of the models. They are also special cases of the more general definitions discussed below.

When M or Y are categorical variables, one option for the regression-based approaches, although only for binary and ordinal variables, is to continue to use linear models (Davis 1980, Hellevik 1984), and another is to formulate models in terms of underlying continuous latent variables and apply linear path analysis to these variables (Heckman 1978, Winship and Mare 1983, Breen, Holm and Karlson 2013). Such methods have been applied also to the analysis of class mobility and education, e.g. by Duncan and Hodge (1963) and Blau and Duncan (1967) using linear path analysis, and by Xie (1989) and Breen and Karlson (2014) using latent-variable formulations. However, these approaches are ultimately unsatisfactory in this application, where latent variables are substantively unappealing and where the categories of at least the class of destination (Y) should be treated as unordered.

More general and flexible methods of mediation analysis than the regression approach have been developed in recent years in the area of formal causal inference (most of it in the potential outcomes framework, of which Imbens and Rubin 2015 give a thorough description). An authoritative overview of this literature is given by VanderWeele (2015). Crucially, causal mediation analysis starts from providing clear definitions of what it means by direct and indirect effects of X on Y (this is typically left undefined in the regression-based methods). The key definition of 'natural effects' was introduced by Robins and Greenland (1992), Pearl (2001), and Robins (2003), and important variants and extensions of it have been proposed by VanderWeele, Vansteelandt and Robins (2014) and VanderWeele and Robinson (2014) (outside the potential outcomes framework, comparable quantities were defined by Didelez, Dawid and Geneletti 2006 and Geneletti 2007); we will discuss these definitions further in Section 3. Such effects can be estimated from observable data, if appropriate assumptions are satisfied. There is now an extensive literature on these assumptions, research designs under which they are more likely to be satisfied, and analysis of the sensitivity of conclusions to violations of them (see VanderWeele 2015 and references therein; because our focus is not causal, these assumptions are not directly relevant to our analysis).

We draw on the ideas and definitions of causal mediation analysis, but translate them to the situation where the parameters of interest are not causal effects but comparable *associations*

in a population. In other words, we want to define and estimate characteristics of a finite population of units which can be used to describe the population in informative ways in terms of the ideas of mediation analysis. These are the relevant and interesting parameters for some substantive research questions, including our questions on education and class mobility among the British populations represented by the three birth cohorts. In particular, we can in this way examine how much of the observed lack of mobility in this period could plausibly be accounted for by the educational differences between social classes that exist among these cohorts, and thus throw light on why educational expansion and reform have had so little effect in equalizing relative mobility chances in Britain over the last half-century.

Our definitions of associations parallel those of causal mediation effects, in essence simply replacing distributions of potential outcomes with conditional distributions of fixed values of the variables in a population. This defines distributions which can be interpreted as conditional distributions of Y given X in standardized populations where some distributions are held constant. An indirect association, which will be our focus, compares groups which have the distributions of M given X that hold in the real population, but share the same reference distribution of X itself. We quantify indirect associations between X and Y by log odds ratios from these standardized distributions. This is a generalisation of the work by Kuha and Goldthorpe (2010) who proposed a special case of this definition with a particular choice of the reference distribution. Outside mediation analysis, the approach is also akin to other 'what-if' statistics which combine true conditional distributions for some variables with reference distributions for others, such as standardized rates in demography (Wachter, 2014), population attributable fraction in epidemiology (Rockhill, Newman and Weinberg, 1998), and 'marginal effects' for illustration of regression results (StataCorp, 2017).

Given a representative sample of data from the population of interest, these associations can be estimated by first specifying and estimating models for the distributions of M given X and of Y given (X, M), and then plugging in fitted values from these distributions in the definitions of the associations. This is equally applicable to variables with ordered or unordered categories (and, with straightforward modifications, also to continuous variables). The resulting estimates are of the same form as estimates of analogous causal effects, and could also be used for that purpose in applications where that was the goal. If the models for M and Y were linear, the estimates would also agree with those of classical regression methods of path analysis. We provide estimates of the standard errors of the estimated associations. When both X and Y are categorical and unordered, the number of distinct odds ratios between them can be large, so we also need tools for summarising their values. In our example we do this by carrying out a second stage of the analysis which describes how the relative sizes of the indirect associations vary by characteristics of the cells (social classes) which define each odds ratio. The methods are described in Section 3 and in an Appendix. Our data are descibed in Section 2 and the analysis of them in Section 4.

In our application, the indirect associations are on average less than half of the corresponding observed (total) associations between origin and destination classes. In other words, less than half of the relative class mobility or immobility between generations that has been observed among these birth cohort populations can be accounted for by existing differences in educational attainment between people from different class backgrounds. This relative contribution of education varies little over time, is slightly larger for women than for men, and varies much more between mobility transitions in different parts of the class structure. It tends to be larger for mobility which spans long hierarchical distances or involves the professional and managerial classes. It is smaller for transitions which involve people staying in the same class as their parents (except for the professional classes), and essentially zero for mobility among classes comprising lower supervisory and technical, semi-routine and routine occupations.

2. Data and variables. We use data from three British birth cohort studies, the MRC National Survey of Health and Development, the National Child Development Study, and the 1970 British Cohort Study. They have followed the life courses of children born in Britain (England, Scotland and Wales) in one week of 1946, 1958 and 1970 respectively (see Wadsworth et al. 2006, Power and Elliott 2006, and Elliott and Shepherd 2006). For the 1958 and 1970 cohorts, the original intended samples included all such births. The 1946 study drew a sample of single births to married women, stratified by the husband's employment; for this cohort, our analysis uses survey weights which allow for the stratification.

The analyses are done separately for each combination of birth cohort and respondent's gender. They involve three variables: a respondent's social class origin, education and social class destination. Our choices and definitions for them follow those of Bukodi et al. (2015) and Bukodi and Goldthorpe (2016), who provide more detailed motivation and information about the variables. For women, we consider only those women who have always worked full time when they have been in employment, thus excluding those who have had periods of part-time employment; for a discussion of this exclusion, see Bukodi et al. (2017).

Social class is categorized using the NS-SEC classification. We use the same seven-class version of it for both origins and destinations, with the classes labelled as shown in Table 1. Destination is defined as a respondent's own class position at the age of 38 years (or when last in employment before then), and origin as the respondent's father's class position when the respondent was aged 10 or 11 (or 15 or 16, if the earlier information is not available). This classification is generally not regarded as fully ordered in terms of more or less advantage, and we will treat the classes as unordered categorical variables.

Education is coded in terms of *relative* education, as defined by Bukodi and Goldthorpe (2016). This begins with a classification of an individual's highest level of qualification achieved by age 37, in eight categories. These are then grouped into four ordered categories, but in a way which is different in different birth cohorts. The groups correspond roughly to the quartiles of levels of qualification in the cohort. The purpose of this is to represent education as a positional good, i.e. that the labour market value of a qualification may depend on its relative position in the current population distribution of qualifications.

Of the original intended cohorts for 1946, 1958 and 1970 respectively, 84%, 81% and 70% of the respondents have at least one of the variables observed, so cohort attrition is fairly small. The number of respondents in our data sets varies from 1020 for women in the 1946 cohort to 7219 for men in 1958. Some of them have missing values in some of the variables. The proportions of missingness are 9–16% for class origin, 0–2% for education, and 23–51% for class destination (40–51% in the 1946 cohort, 23–34% in 1958 and 1970). The missing values have been multiply imputed to allow for the inclusion of the incomplete observations in the analysis. The imputation, which is based on MCMC estimation of a saturated model for the joint distribution of the variables, is described in Appendix A of Bukodi, Goldthorpe and Kuha (2017). Ten multiply imputed datasets were used for our analyses, as explained further in Section 3.4.

The estimated marginal distributions of the class variables for men and women in each cohort are shown in Tables 1 and 2. There have been marked changes in these class distributions over time, in particular in that the proportions of people in the lower-numbered ('white-collar' and salaried) classes have increased over time, while those in the higher-numbered ('blue-collar' and wage-earning) ones have decreased.

Relative class mobility is, however, not described by changes in these univariate marginal distributions, but by associations in the mobility table between individuals' classes of origin and destination. Here this is a 7×7 contingency table for each gender and cohort (an example is shown in Table 3 in Section 4). As a measure of the associations, we use the log odds ratios (log ORs) for different 2×2 subtables of the mobility table, each defined by the intersection

TABLE 1

Estimated distributions (in %) of class origins ('Orig.') and class destinations ('Destin.') among the members of the 1946, 1958 and 1970 birth cohorts, for men.

	1946 cohort		1958 cohort		1970 cohort	
Class	Orig.	Destin.	Orig.	Destin.	Orig.	Destin.
1: Higher managers and professionals	4.5	11.8	6.7	16.2	11.3	21.4
2: Lower managers and professionals	9.6	26.1	15.3	20.5	17.4	21.4
3: Intermediate occupations	9.8	10.1	14.7	9.2	7.4	9.7
4: Small employers and own account workers	10.0	10.6	5.7	14.4	14.1	14.4
5: Lower supervisory and technical occupations	12.6	12.5	19.3	11.2	13.9	8.8
6: Semiroutine occupations	16.6	12.8	10.9	12.9	14.0	12.8
7: Routine occupations	37.0	16.2	27.3	15.6	21.7	11.6
Total	100	100	100	100	100	100
n	2394		7219		5979	
(n_{obs})	(2078)	(1175)	(6557)	(5582)	(5022)	(4124)

n denotes the total number of respondents, including those for whom the variable is not observed.
n_{obs} denotes the number of respondents for whom the variable is observed.
The percentages in the table (and the results for all other tables and figures of this paper) were estimated using 10 multiply imputed datasets to allow for the missing data.

TABLE 2

Estimated distributions (in %) of class origins ('Orig.') and class destinations ('Destin.') among the members of the 1946, 1958 and 1970 birth cohorts, for women.

	1946 cohort		1958 cohort		1970 cohort	
Class	Orig.	Destin.	Orig.	Destin.	Orig.	Destin.
1: Higher managers and professionals	3.9	2.3	6.4	7.4	11.6	12.9
2: Lower managers and professionals	7.8	19.8	17.4	23.5	19.2	27.6
3: Intermediate occupations	9.0	34.5	14.4	28.0	6.9	28.6
4: Small employers and own account workers	8.5	6.7	5.0	6.7	13.0	5.8
5: Lower supervisory and technical occupations	15.3	2.5	18.1	1.5	13.9	1.4
6: Semiroutine occupations	19.7	17.0	9.9	19.0	13.9	16.2
7: Routine occupations	35.8	17.2	28.8	13.9	21.6	7.6
Total	100	100	100	100	100	100
n	1020		3535		2432	
(n_{obs})	(913)	(614)	(3162)	(2372)	(2035)	(1613)

n denotes the total number of respondents, including those for whom the variable is not observed.
n_{obs} denotes the number of respondents for whom the variable is observed.

of two origin classes (rows) and two destination classes (columns). There are 441 such log ORs, and we may want to use any of them to summarise different aspects of class mobility. Their values are mostly different from 1 (and often quite far from it), indicating substantial immobility (lack of fluidity) of social class between generations among these birth cohorts. Patterns and trends in these bivariate associations have been analysed in previous literature, as discussed in Section 1. Our aim in this article is to further examine how they may be mediated by education. We will return to this analysis for the birth cohort data in Section 4, after a general method for doing it has been described in Section 3.

3. Mediation analysis for associations.

3.1. *Total effects and associations.* Consider variables X, M and Y as represented in Figure 1, where M is a mediator between X and Y. The models may also be conditional on observed confounders, but these are omitted from the notation here. We take all three variables to be categorical because this is the situation in our application, but the ideas discussed here apply also when they are continuous variables (the case where M is continuous is considered in Appendix A.3). The numbers of distinct values of X, M and Y are denoted by J, K and L respectively. Suppose that we are interested in a finite population of N units and that we observe (X_i, M_i, Y_i) for a sample of $n \leq N$ units i from that population. In our application, X is an individual's class of origin, M their relative education, and Y their class of destination, with $J = L = 7$ and $K = 4$, and we are interested in six distinct populations of British adults, for men and women in each of the three birth cohorts. Marginal and conditional distributions of variables in the population are denoted by $p(\cdot)$ and $p(\cdot|\cdot)$ respectively.

Two kinds of estimands could be of interest here: *causal effects* of X on Y, or *associations* between X and Y in a population. Our only goal will be to estimate associations. However, because our definitions are strongly motivated by the causal ones, and because the similarities and differences between the two kinds of parameters are illuminating on both of them, we start by discussing associations and causal effects in parallel. In this section we will often refer to both of them as 'effects'. We discuss first *total effects*, and then in Section 3.2 the central concept of *indirect effects* which also involve M (corresponding *direct effects* are considered in Section 3.3). We focus first on their definitions and interpretations, before estimation of the associations is described in Section 3.4 and Appendix A.

The two kinds of target parameters involve different conceptions of what values the 'population' consists of. For associations, these are (X_i, M_i, Y_i), treated as fixed values for each unit $i = 1, \ldots, N$. This population could be observed in full if we carried out a census of it. Causal effects, in contrast, are defined in terms of two kinds of potential outcomes: $M_i(x)$, the value that M would have for unit i if X was set to the value x for that unit, and $Y_i(x, m)$, the value that Y would have for i if (X, M) were set to (x, m). The population values are then $M_i(x)$ and $Y_i(x, m)$ for all possible values of x and m, for the units $i = 1, \ldots, N$. Here this population of potential outcomes would consist of J sets of values of $M_i(x)$ and $J \times L$ sets of $Y_i(x, m)$ for each i. Only one of the $Y_i(x, m)$ and one of the $M_i(x)$ can ever be observed for any i, even for the n units in a sample.

All of the effects are defined in terms of comparisons between two distinct values of X at a time. We denote them by $X = r$ and $X = s$. Many different pairs (r, s) may be of interest and there need not be any one value of X which is always treated as the reference value. In our application, where X has seven levels, there are 21 distinct pairs (r, s).

Consider first causal total effects in this context. For a unit i, the effect on Y of X being set to $X = s$ rather than $X = r$ is defined as a comparison between $Y_i(r, M_i(r))$ and $Y_i(s, M_i(s))$. In this, M is not set independently but assumes the value it will naturally have when X is set to r or s. This matches the intuitive idea that a total effect should incorporate all effects of X on Y, including those that arise from the effect of X on any M and the consequent effect of M on Y. We can then also write $Y_i(x, M_i(x)) = Y_i(x)$.

Instead of unit-level causal effects, we can only estimate their aggregates in a population of units. These are defined as comparisons of the distributions $p(Y_i(r))$ and $p(Y_i(s))$ over $i = 1, \ldots, N$. Denoting $\pi^*_y(x) = p(Y_i(x) = y)$, different comparisons of these proportions may be considered as the parameters which quantify a total effect, for example the differences $\pi^*_y(s) - \pi^*_y(r)$. We focus on the log odds ratios (log ORs)

$$\psi^{TE}_{rs.tu} = \log \frac{\pi^*_u(s)/\pi^*_t(s)}{\pi^*_u(r)/\pi^*_t(r)} \tag{1}$$

for all $r \neq s = 1, \ldots, J$ and $t \neq u = 1, \ldots, L$. These are the log ORs for the 2×2 subtables in the $J \times L$ table which cross-classifies x by $Y_i(x)$ for $i = 1, \ldots, N$ (i.e. a table where the counts on each row sum to N).

To define associational total effects, we refer instead to the population defined as (X_i, M_i, Y_i) for $i = 1, \ldots, N$. Consider the cross-tabulation of X_i by Y_i in it (i.e. a table where the counts over the whole table sum to N), and the conditional probabilities $\pi_y(x) = p(Y_i = y | X_i = x)$ in this table. The log ORs of interest are then

$$\theta^{TE}_{rs.tu} = \log \frac{\pi_u(s)/\pi_t(s)}{\pi_u(r)/\pi_t(r)}. \tag{2}$$

Although the causal effects $\psi^{TE}_{rs.tu}$ and the associations $\theta^{TE}_{rs.tu}$ are quite different parameters, they are *estimated* by sample quantities of the same kind. Letting $\hat{\pi}_y(x) = \hat{p}(Y = y | X = x)$ denote estimates of conditional probabilities derived from the observed (Y_i, X_i) for $i = 1, \ldots, n$, the estimate is

$$\hat{\theta}^{TE}_{rs.tu} = \log \frac{\hat{\pi}_u(s)/\hat{\pi}_t(s)}{\hat{\pi}_u(r)/\hat{\pi}_t(r)} \tag{3}$$

for both parameters, but under different assumptions about the observed data. Estimation of associations relies on the assumption (call it S) that the observed units are a representative sample from the units in the population, so that estimates from the sample (possibly with sampling weights) can be generalised to this population. This assumption is most convincingly satisfied if the data are a probability sample from the population. For estimation of a causal effect we need instead the standard assumptions of causal inference (call them C), the most prominent of which is that there should be no unmeasured confounding of the effect of X on Y. This may be thought of as an assumption of representative sampling of the potential outcomes. It is most convincingly satisfied when the values of X were randomized to the units in the sample. If assumption C holds, $\hat{\theta}^{TE}_{rs.tu}$ is an estimate of the causal log OR $\psi^{TE}_{rs.tu}$ among the n units in the sample, and if S also holds it is also an estimate of $\psi^{TE}_{rs.tu}$ among the N units in a larger population. If neither assumption holds, $\hat{\theta}^{TE}_{rs.tu}$ is just an estimate of itself, i.e. a descriptive statistic for the sample. If assumption S holds (even if C does not), $\hat{\theta}^{TE}_{rs.tu}$ is an estimate of the population association $\theta^{TE}_{rs.tu}$. This last case is the goal of our analysis.

3.2. *Indirect effects and associations.* Definitions of indirect and direct effects break the link between X and the mediator M, and vary only one of them. For example, a unit-level *indirect causal effect* of setting $X = r$ rather than $X = s$ is a comparison between $Y_i(x^*, M_i(r))$ and $Y_i(x^*, M_i(s))$, where x^* is a given fixed value. This captures the idea that an indirect effect is the change in Y when M changes as if in response to a change in X, but X itself does not change but remains fixed at x^*.

The parameters which we can aim to estimate are again defined not for individual units but for distributions over populations of units. We may consider four kinds of distributions:

$$p(Y_i(x^*, M_i(x))), \tag{4}$$

$$\sum_m p(Y_i(x^*, m))\, p(M_i(x) = m), \tag{5}$$

$$\sum_m p(Y_i(m) | X_i = x^*)\, p(M_i = m | X_i = x), \quad \text{and} \tag{6}$$

$$\sum_m p(Y_i | X_i = x^*, M_i = m)\, p(M_i = m | X_i = x). \tag{7}$$

For each of them, an indirect effect is defined as a comparison of these quantities with $x = r$ against $x = s$, for some fixed x^*. The definitions (4), (5) and (6) imply three different causal indirect effects. We add the population distribution in (7), which leads to the purely associational definition which will be our focus.

The causal quantities (4)–(6) reflect the idea of an indirect effect in slightly different ways. In (4), the crucial feature is that the potential outcomes for M and for Y refer to same unit i, i.e. we consider the value of Y for i if X had been set to x^* and M had been set to the value it would have for i if X were set to x. When $x^* = r$ or $x^* = s$, this defines the *natural indirect effect* of X on Y. This setting is relaxed in (5), which is the distribution of $Y_i(x^*, M)$ over i when M is set not to each unit's own potential outcome $M_i(x)$ but to a value drawn from the distribution of $M_i(x)$ across all the units. VanderWeele, Vansteelandt and Robins (2014) call the resulting effects the *interventional indirect effects*. In (6), furthermore, (X_i, M_i) are treated as fixed values for the units, but it is also conceived that each unit could have had a different value of M. The only potential outcomes considered are then $Y_i(m)$ given different settings of M, and (6) is the distribution of $Y_i(M)$ for units i for whom $X_i = x^*$ when M is drawn from its distribution among the units for whom $X_i = x$. This kind of effect was introduced by VanderWeele and Robinson (2014). Finally, (7) involves no potential outcomes but only distributions of fixed (X_i, M_i, Y_i) in the population, as discussed below.

The quantities (4)–(7) can again be estimated, under different assumptions, by the same sample quantity, namely $\sum_m \hat{p}(Y = y|X = x^*, M = m)\,\hat{p}(M = m|X = x)$ where the $\hat{p}(\cdot|\cdot)$ are estimates of these conditional distributions (we will discuss the estimation further in Section 3.4). Generalisation to a population of $N > n$ units again requires that the observed units are a representative sample from this population. Only this assumption is relevant to (7) and to our analysis. Estimating the causal quantities (4)–(6), on the other hand, again requires assumptions about representativeness of the potential outcomes that are observed in the sample, in a form which is decreasingly demanding going from (4) to (6). For (4), we need the assumption that there is no unmeasured confounding of the effect of X on M and on Y, or of the effect of M on Y, plus the 'cross-world independence assumption' that the potential outcomes $Y_i(x^*, m)$ and $M_i(x^\dagger)$ are independent for $x^* \neq x^\dagger$ (see VanderWeele 2015, S. 7.3). The cross-world independence assumption can be omitted for (5) and (6), and for (6) the conditions which refer to effects of X can also be omitted.

Having now motivated the associational quantity (7) partly by reference to corresponding causal quantities, we will from here on focus solely on such associations. Before we do that, however, we make one further modification to the parameter of interest. This is needed because (7) is defined with a single value in the role of x^*. This is not ideal in our application, where no one x^* is natural for all units and for all the comparisons that we will want to consider. To allow for this, we take x^* to be not a single fixed value but drawn from a fixed distribution $p_0(X)$, which we call the *reference distribution* of X. Averaging (7) over it gives

$$\begin{aligned}\pi_y^{IE}(x) &= \sum_{x^*}\sum_m p(Y = y|X = x^*, M = m)\, p(M = m|X = x)\, p_0(X = x^*) \qquad (8)\\ &= \sum_m p_0(Y = y|M = m)\, p(M = m|X = x),\end{aligned}$$

where $p_0(Y = y|M = m) = \sum_{x^*} p(Y = y|X = x^*, M = m)\, p_0(X = x^*)$. This would reduce back to (7) if we chose $p_0(X = x^*) = 1$ for a single value x^*. We then define

$$\theta_{rs.tu}^{IE} = \log \frac{\pi_u^{IE}(s)/\pi_t^{IE}(s)}{\pi_u^{IE}(r)/\pi_t^{IE}(r)} \qquad (9)$$

for any $r \neq s = 1, \ldots, J$ and $t \neq u = 1, \ldots, L$, and with all four probabilities on the right-hand side of (9) calculated using the same reference distribution $p_0(X)$. The log odds ratios (9) are the *indirect associations* which we will focus on for the rest of this article. They

are a generalisation of a definition first proposed by Kuha and Goldthorpe (2010), who considered associations which are equal to (8)–(9) for a particular choice of $p_0(X)$, as discussed below.

These indirect associations can be interpreted as associations in certain standardized, synthetic populations. To motivate this, consider first the true conditional probabilities of Y, which determine the total associations, written in the slightly unusual form

$$\begin{aligned}\pi_y(x) &= \sum_m p(Y=y|X=x,M=m)\, p(M=m|X=x) \\ &= \sum_{x^*}\sum_m p(Y=y|X=x^*,M=m)\, p(M=m|X=x)\, p^*_x(X=x^*) \qquad (10)\end{aligned}$$

where $p^*_x(X=x)=1$ denotes a distribution with only one value x. This highlights the fact that in each of the J groups defined by $X=1,\ldots,J=x$ in the actual population, every unit has $X=x$ and M follows its conditional distribution given that $X=x$. In contrast, in the probabilities $\pi^{IE}_y(x)$, as defined in (8), the conditional distributions of Y and M are the same population distributions as in (10), but the distribution of X is changed from $p^*_x(X)$ to the reference distribution $p_0(X)$. This defines standardized groups which differ from each other in the distribution of M in the same way as do the groups with different values of $X=x$ in the real population, but which have the same (reference) distribution of X itself. This captures the basic idea that indirect associations should compare situations where M varies as if in response to differences in X, but (the distribution of) X itself is held fixed; the latter ('direct') contribution of X can be viewed as representing factors other than M which are also associated with X and with Y. In our application, the indirect log ORs compare the log odds of different classes of destination (Y) between hypothetical groups which are like the actual origin classes (X) in the population in their distributions of educational attainment (M), but which each have the same distribution $p_0(X)$ of other aspects of class of origin which are associated with social mobility.

The exact values of these associations depend to some extent on the reference distribution $p_0(X)$. How then should it be selected? The simplest choice for a specific log OR $\theta^{IE}_{rs.tu}$ would be to fix X at r or s for everyone, i.e. to choose $p_0(X=r)=1$ or $p_0(X=s)=1$. This is most suitable when there is a natural baseline level for each comparison, for example if X is ordinal and we use the lower level as the baseline. If not, an alternative is to give equal weight to both r and s, i.e. $p_0(X=r)=p_0(X=s)=1/2$; this was the reference distribution proposed by Kuha and Goldthorpe (2010). With both of these choices, the reference distribution for the probabilities $\pi^{IE}_x(y)$ which define a $\theta^{IE}_{rs.tu}$ depends on the rows r,s, and thus each $\pi^{IE}_x(y)$ will have different values in different log ORs for the same table. This is a disadvantage, because it compromises the standardization interpretation discussed above. In particular, it is then not possible to represent the standardized population in the form of a single contingency table of the $\pi^{IE}_x(y)$. Our preference is to use instead a reference distribution $p_0(X)$ which is the same for all the $\theta^{IE}_{rs.tu}$ in the same table (an example of a table of $\pi^{IE}_x(y)$ from such a distribution is shown in Table 3 in Section 4). A potential disadvantage of such a choice is that for a given $\theta^{IE}_{rs.tu}$, the calculation of $\pi^{IE}_x(y)$ from (8) will then involve averaging $p(Y|X=x^*,M)$ — in essence, averaging the 'direct effects' of X — also over values x^* which are not r or s, which may be substantively less appealing especially if there are strong interactions between X and M in this distribution for Y. The reference distribution can be chosen by the researcher, in a way which is judged to be appropriate for the application. In our analysis of social mobility we have set it, for all log ORs, to the estimated marginal distribution $\hat{p}(X)$ of classes of origin in the population, as discussed further in Section 4. In that analysis, the specific choice of the reference distribution does not have a substantial effect on the conclusions.

3.3. *Direct effects and effect decompositions.* Mirroring the logic of indirect effects, *direct effects* compare distributions of Y when X assumes different values but M is not allowed to change accordingly. Paralleling (8) and (10), we can define

$$\pi_y^{DE}(x) = \sum_{x^*} \sum_{m} p(Y = y|X = x^*, M = m)\, p_0(M = m)\, p_x^*(X = x^*) \tag{11}$$

where again $p_x^*(X = x) = 1$, and $p_0(M = m)$ is a reference distribution for M (it could be derived, for example, as $p_0(M = m) = \sum_{x^\dagger} p(M = m|X = x^\dagger)\, p_0(X = x^\dagger)$ where $p_0(X)$ is a reference distribution for X). The probabilities $\pi_y^{DE}(x)$ can be used to define *direct associations* as the log odds ratios $\theta_{rs.tu}^{DE}$, analogously with $\theta_{rs.tu}^{IE}$ in (9). They can again be interpreted as comparisons of the distributions of Y between two standardized groups. These have $X = r$ for every unit in one group and $X = s$ in the other, as in a total association, but now M has the same reference distribution $p_0(M)$ in both groups. In our analysis, this would mean that we compare the log odds of different classes of destination between groups which are like classes of origin r and s in other respects, but which have the same distribution $p_0(M)$ of educational attainment.

A common aim of mediation analysis is to 'decompose' a measure of a total effect into the sum of direct and indirect effects. Deviating for the moment from our focus on log ORs, suppose that the total association was quantified by the difference $\pi_y(s) - \pi_y(r)$. When plugged into (8) and (10), some choices of the reference distribution $p_0(X)$ yield the exact decomposition $\pi_y(s) - \pi_y(r) = (\pi_y^{IE}(s) - \pi_y^{IE}(r)) + (\pi_y^{DE}(s) - \pi_y^{DE}(r))$. This is achieved if $p_0(X = r) = p_0(X = s) = 1/2$ (as in Kuha and Goldthorpe 2010), or if $p_0(X = r) = 1$ for $\pi_y^{IE}(r)$ and $\pi_y^{IE}(s)$ and $p_0(X = s) = 1$ for $\pi_y^{DE}(r)$ and $\pi_y^{DE}(s)$ (or vice versa), thus using different reference distributions for the two associations; the latter is also the basis of the corresponding decomposition for the natural causal indirect and direct effects (see e.g. VanderWeele 2015, A.2.1). Further instances of decompositions are possible under specific parametric models for Y and M. For categorical variables, these can be obtained by specifying linear models for them or for hypothetical underlying continuous latent variables (Winship and Mare 1983; Breen, Holm and Karlson 2013; Breen and Karlson 2014), or as approximations for non-linear models (see VanderWeele 2015).

We will not, however, make use of such decompositions, because for our purposes they are not necessary or even helpful. This is because the standardized scenarios which provide the interpretations of direct and indirect associations are not inherently paired. Each of these log ORs implies a comparison of two groups with different joint distributions of X and M. Given $X = r$, this distribution is $p_0(X)p(M|X = r)$ for the indirect $\theta_{rs.tu}^{IE}$, $p_r^*(X = r)p_0(M)$ for the direct $\theta_{rs.tu}^{DE}$, and $p_r^*(X = r)p(M|X = r)$ for the total $\theta_{rs.tu}^{TE}$. There is no very convincing sense in which the first two of these are a matched pair which together match up with the third, as would be implied by the expectation that the corresponding direct and indirect associations should add up to the total one. Instead, we will focus on one kind of association — in our application, the indirect one — and view the difference between it and the total association as a residual of the total which is not accounted for by the mediator M, rather than as a specific well-defined direct effect.

We will nevertheless want to assess the relative magnitudes of different indirect associations. We will do this by comparing them to the corresponding total associations, specifically by considering the ratios $\theta_{rs.tu}^{IE}/\theta_{rs.tu}^{TE}$. In our application this can be interpreted as the ratio of the log OR for an origin-destination association in a standardized population where people from different classes of origin differ *only* in their educational attainment ($\theta_{rs.tu}^{IE}$), against the log OR in the actual population where they differ not only in education but also in other characterics which are associated with class of destination ($\theta_{rs.tu}^{TE}$). This ratio can also be negative or greater than one because the indirect association can be stronger, or have a different sign, than the total association (although this turns out to be rare in our data).

3.4. *Estimation of the associations.* As can be seen from the definitions of $\pi_y^{IE}(x)$ and $\pi_y^{DE}(x)$ in (8) and (11), these probabilities, and so also the log odds ratios $\theta_{rs.tu}^{IE}$ and $\theta_{rs.tu}^{DE}$, are functions of the population probabilities $p(M = z|X = x)$ and $p(Y = y|X = x, M = z)$. They can be estimated by plugging in sample estimates of these conditional probabilities (this is also true for $\pi_y(x)$ and the total log ORs $\theta_{rs.tu}^{TE}$, as shown in (10), but they can also be estimated without involving M). This can be done for all combinations of x, m and y at once by using matrix formulations, as shown in Appendix A.1. Variance matrices of the estimated log ORs, and of quantities such as their ratios, can be derived with the delta method, as described in Appendix A.2. Bootstrap methods of variance estimation could also be used.

All that then remains in any specific application is to estimate the probabilities, from models specified for $p(M|X)$ and $p(Y|X, M)$. In our analysis we have used saturated models for them. This implies, in particular, that the model for Y includes an interaction between X and M. An alternative would be to use non-saturated models, for example a multinomial logistic model for Y without the interaction. Howsoever these conditional probabilities are specified, the rest of the estimation of the associations remains the same.

It is not a new or surprising conclusion that estimation in mediation analysis starts with estimation of conditional distributions for M and Y. As discussed in Sections 3.1 and 3.2, estimates of effects in causal mediation analysis are also based on these distributions, although they are then treated as estimates of distributions of potential outcomes (for such estimates with a causal focus, see e.g. Pearl 2001, Imai, Keele and Tingley 2010, Loeys et al. 2013, and other examples in VanderWeele 2015). This means that our estimates of indirect and direct associations could also be used to estimate analogous causal effects, in applications where that was the goal. This could be useful, in particular, for the case of unordered polytomous outcomes Y, which has been somewhat less discussed in the causal mediation literature. Similarly, those estimates in regression-based mediation analysis which are expressed in terms of regression coefficients can typically also be derived from these general expressions of conditional distributions, in the very special cases where that is exactly or approximately possible.

The data in our application had been multiply imputed to allow for missing data. Estimates of the conditional probabilities were first calculated separately for each of the multiply imputed datasets (in the case of the 1946 cohort, using also the survey weights), and then combined using standard multiple imputation methods to get estimates $\hat{p}(Y|X, M)$ and $\hat{p}(M|X)$ and their variance matrices. These then served as the starting point for the estimation of the log ORs, using the calculations described in Appendices A.1 and A.2.

We have assumed that the mediating variable M is categorical, because that is how it is treated in our application. More generally, M could also be continuous even when X and Y are categorical (and the associations of interest thus remain log ORs). The sums over M above then become integrals, which need to be evaluated as part of the estimation process. In Appendix A.3 we sketch one way of doing this, using Monte Carlo integration.

4. Analysis of social class mobility and education. We now return to the analysis of social class mobility among the British birth cohorts. Here the variables X, M and Y are a person's class origin, relative education and class destination respectively. Our goal is to assess how much of the total associations between origins and destinations may be accounted for by differences in educational attainment between people from different class origins, and how this varies between cohorts, between genders, and between transitions in different parts of the class structure. To examine this, we estimated the total and indirect associations described in Section 3, for each of the 441 log odds ratios between different origin and destination classes, for men and for women, in each of the 1946, 1958 and 1970 cohorts.

The birth cohort samples were censuses or probability samples of births in a single week, and we can reasonably treat them as representative also of people born in Britain in those

TABLE 3

Estimated conditional probabilities of destination class (y) given origin class (x), for men in the 1970 birth cohort. The probabilities on the left are estimates of the actual probabilities $\pi_y(x)$ in the population, and those on the right are estimates of the 'indirect' probabilities $\pi_y^{IE}(x)$ where origin classes differ only in their distributions of education.

	Destination													
	Total $[\hat{\pi}_y(x)]$							Indirect $[\hat{\pi}_y^{IE}(x)]$						
Origin*	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	.42	.26	.10	.09	.04	.06	.03	.30	.25	.10	.12	.07	.09	.07
2	.30	.29	.11	.10	.07	.09	.05	.26	.23	.10	.13	.08	.11	.09
3	.23	.24	.10	.16	.06	.10	.12	.22	.22	.10	.14	.09	.12	.10
4	.17	.19	.09	.23	.07	.13	.12	.21	.21	.10	.15	.09	.13	.12
5	.20	.18	.09	.15	.14	.15	.09	.19	.21	.10	.15	.09	.13	.12
6	.15	.19	.09	.14	.11	.19	.13	.18	.20	.10	.15	.09	.14	.13
7	.11	.17	.09	.15	.10	.16	.22	.17	.20	.10	.16	.09	.15	.13

∗ See Table 1 for the labels of the classes.

years. The samples have since been reduced by cohort attrition, but at 20–30% this nonresponse rate is relatively small. Item nonresponse in individual variables has been addressed through multiple imputation, under the assumption that these data are missing at random.

As the reference distribution $p_0(X)$ of class origin, for all log ORs for both men and women in a given cohort, we use the marginal distribution $\hat{p}(X)$ of origins estimated from data for all the respondents in that cohort. We have also repeated the analysis using a reference distribution which is selected separately for each log OR $\theta_{rs.tu}^{IE}$ to match its origin classes, with equal probabilities $p_0(X = r) = p_0(X = s) = 1/2$. The estimates from this alternative analysis are shown in supplementary materials to this article, in the same form as Figure 2 and Tables 4 and 5 below (the right-hand side of Table 3 is not available in this case, as noted in the discussion at the end of Section 3.2). This change of the reference distribution leaves the results discussed here essentially unchanged.

The starting point of the analysis is the set of estimated probabilities of destinations given origins. Table 3 shows them for men in the 1970 cohort, showing estimates of both the actual population probabilities $\pi_y(x)$ and the indirect probabilities $\pi_y^{IE}(x)$. The estimated total and indirect log ORs are calculated from these probabilities. For example, the four cells in the corners of the table define a 2×2 table where the two origin classes and the two destination classes are both classes 1 and 7. The total log OR for it is 3.29. This indicates substantial barriers to mobility between these classes, so that when a man's father was in Class 1 (higher managers and professionals), the man himself was much more likely to end up in this class rather than in Class 7 (routine occupations) — and vice versa for men whose fathers were in Class 7. The indirect log OR for this is 1.22. It shows that some of the total association can be accounted for by the fact that men from origin Class 1 are relatively more likely than men from Class 7 to attain those levels of education which are associated with higher probabilities of ending up in destination Class 1 rather than 7. The ratio of these indirect and total log ORs, however, is only 0.37, so educational differences alone account for only 37% of the total association (lack of mobility) between these classes.

Sociologically, there is every reason to expect that the relative contribution of education to class mobility will be different for different kinds of class transitions. This raises the practical challenge of how to examine the results across the many log ORs which may be calculated for the mobility table. Here we do this, first, by reporting the full results for an interesting subset of the log ORs. We then analyse a larger subset of them, focusing on the ratios between indirect and total associations. We first compare the average levels of these ratios between

men and women and between the cohorts, and then also fit descriptive regression models which reveal how the ratios vary between different kinds of mobility transitions.

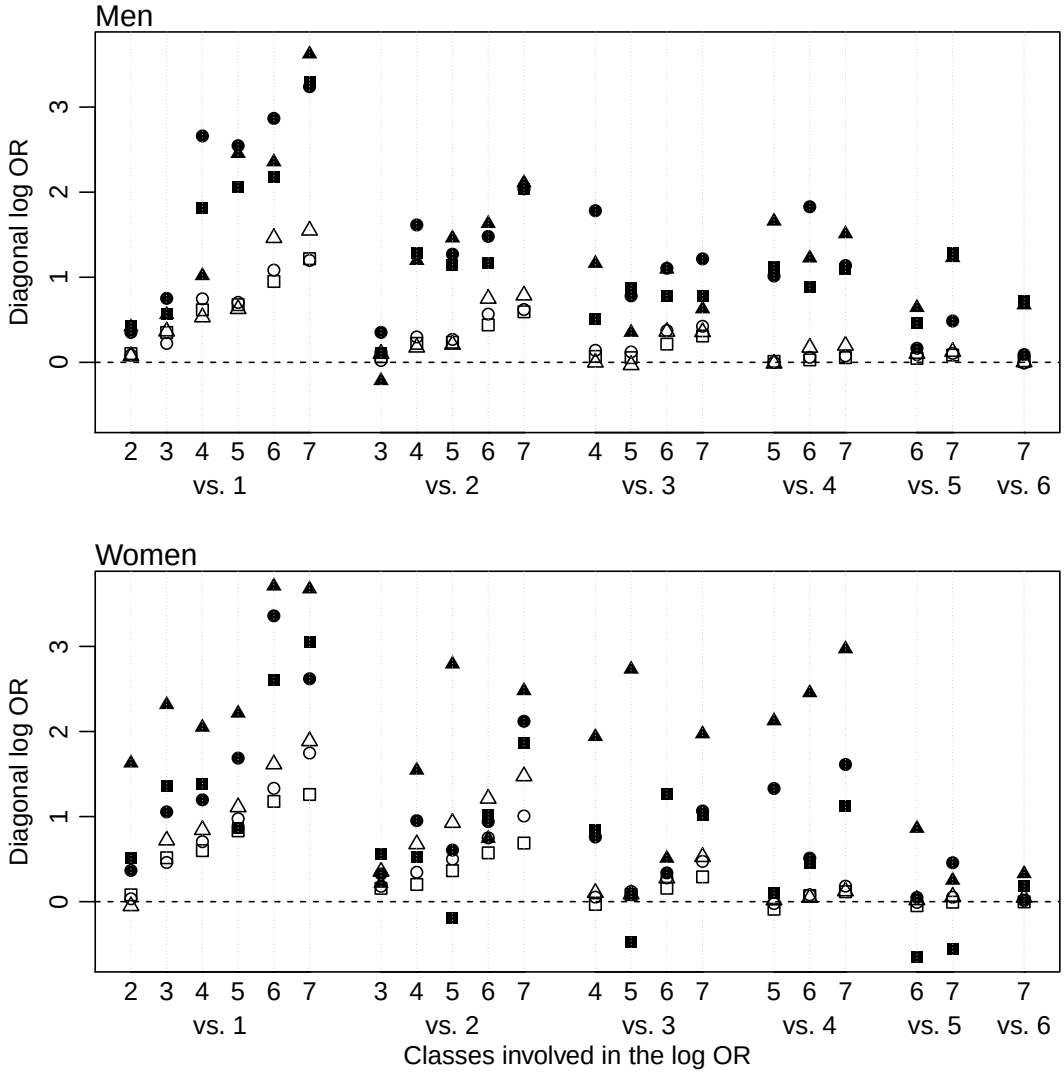
Figure 2 shows the estimated total and indirect log ORs for men and women in all three cohorts, for the 21 'diagonal' log ORs where the two origin classes are the same as the two destination classes. For example, the two log ORs discussed above for illustration are shown in the sixth column ('7 vs. 1') of the upper plot, the total association (3.29) by the filled-in square and the indirect association (1.22) by the open square. Some regularities may be observed already here. The total associations are larger when they involve a large hierarchical distance between the classes (roughly, the difference between their numbers). This is also true for the indirect associations, in cases which involve the professional and managerial Classes 1 and 2 (and to a lesser extent the intermediate Class 3). For these transitions, the ratios between indirect and total log ORs are commonly between 0.2 and 0.5. In contrast, for transitions among Classes 4–7 (the small employers, the self-employed, and the 'blue-collar' occupations), the indirect associations (the open symbols) are consistently close to zero, so educational differences appear to account for very little in these cases (for men in the 1970 cohort this can also be seen in Table 3, where the probabilities $\hat{\pi}_x^{IE}(y)$ are very similar in all of Classes 4–7). There are no obvious regularities in the differences between the cohorts. The total associations are often somewhat smaller for women than for men, resulting in higher ratios of indirect to total associations for women.

Here we could continue with a more comprehensive analysis of the indirect log ORs themselves, examining patterns in them across different genders, cohorts and class transitions. This would mean carrying out the kinds of analysis that, for example, Bukodi et al. (2015, 2017) did for the total associations, but now with the indirect $\hat{\pi}_y^{IE}(x)$ as the starting point. However, the main substantive research question, which is how much of the observed (im)mobility of social class could be accounted for by class differences in education, directs the focus instead on the sizes of the indirect associations relative to the corresponding total ones. In the rest of this section we examine this for the ratios between the estimated indirect and total log ORs, using them as a summary of the relative strengths of the indirect ones.

The analysis of these ratios is initially complicated by the cases where the estimated total log OR is itself very small, so that the ratio takes extreme values which obscure any patterns among the set of ratios as a whole. This may happen when the sampling variation in the estimates is high (for less common combinations of origin and destination) or when the total population association is close to zero. To remove this distraction, we make the purely ad hoc choice of limiting these summary analyses to those cases where the total association is not very small and is estimated fairly precisely. As a cut-off, we include only those ratios where the (Wald test) p-value of the total log OR is less than 0.10. We also omit the small number of cases (at most 18) where the total log OR is negative. For them, the indirect association is often still positive and the ratio is thus negative, which somewhat obscures the patterns among the bulk of the ratios which are positive. Many of these log ORs correspond to mobility transitions where individuals stay in the same class as their fathers and where such 'inheritance' effects work in the opposite direction from differences in education; many of these cases involve the self-employed Class 4 where inheritance is especially strong. Such cases are best considered separately.

These exclusions leave us with between 65 (for women in 1946) and 260 (for men in 1958) of the 441 possible ratios. The top part of Table 4 shows summary statistics of them, separately for each combination of cohort and gender. The average ratios are between 0.30 and 0.49. They are higher for women than for men, but there are no very clear trends across the cohorts. The bottom part of Table 4 then summarises these ratios separately by cohort and by gender. For example, the analysis by cohort starts from the $882 = 441 + 441$ ratios for men and for women in each cohort, and then includes the 145 of these where the estimated total

FIG 2. *Estimates of diagonal log odds ratios between class of origin and class of destination, for men and women in the three birth cohorts. Here open symbols denote the indirect log ORs, △ for 1946, ◯ for 1958 and □ for 1970, and the filled-in versions of the same symbols denote the corresponding total log ORs. For example, the diagonal log ORs involving classes 1 and 7 for men in the 1970 cohort, which are discussed in the text for illustration, are shown in the sixth column ('7 vs. 1') of the upper plot, the total log OR (3.29) by the filled-in square and the indirect one (1.22) by the open square.*



association has $p < 0.10$ in every cohort. Here the average ratio is again larger for women (0.46) than for men (0.36), and the 95% confidence interval for their difference is (0.01; 0.20). The average decreases over time, but less clearly; the difference between the 1970 and 1946 cohorts is -0.06, with a 95% confidence interval of $(-0.26; +0.14)$.

Finally, in Table 5 we examine how the ratios vary between different kinds of class transitions. This is described by linear models where the response variable is the estimated ratio $\hat{\theta}^{IE}_{rs.tu}/\hat{\theta}^{TE}_{rs.tu}$. The sets of ratios which are used as data for these models are the same ones which were included at the top of Table 4. The units of analysis are then in effect the 2×2 tables for which the log ORs are calculated, each of which is defined by two origin classes $r < s$ and two destination classes $t < u$. The explanatory variables of the models are characteristics of these classes, or of the log ORs themselves. The same explanatory variables are

TABLE 4

Summary statistics of the estimated ratios between indirect and total log odds ratios between origin and destination classes, for subsets of the log ORs among men and women in the three birth cohorts.

By combinations of cohort and gender:

	1946 cohort		1958 cohort		1970 cohort	
	Men	Women	Men	Women	Men	Women
Number of ratios*	126	65	260	161	227	122
Mean ratio	.35	.43	.30	.49	.31	.40
(5%-95%)‡	(.01–.71)	(.04–.88)	(-.01–.58)	(.10–.95)	(.03–.58)	(.12–.74)

Separately by cohort and by gender:

	Cohort:				
	1946	1958	1970	Men	Women
Number of ratios†	145	145	145	277	277
Mean ratio	.42	.38	.36	.36	.46
(5%-95%)‡	(.05–.82)	(.08–.71)	(.06–.60)	(.08–.59)	(.11–.80)

∗ Number of log ORs for which $p < 0.10$ for the estimated total association
† Number of log ORs for which $p < 0.10$ for the estimated total association for all cohorts (145) or for both men and women (277).
‡ The range between 5th and 95th percentiles of the estimated ratios.

used in every cohort-gender model. They were identified through exploratory model selection and substantive considerations, in particular by paralleling some of the specifications that were used by Bukodi, Goldthorpe and Kuha (2017) in their 'topological' log-linear models for the total associations in these same data.

The first explanatory variable is the total log OR itself. The second, labelled 'White-collar to/from Other' in Table 5, is 1 if $r, t \leq 3$ and $s, u > 3$, and 0 otherwise. It is an indicator for associations which correspond to mobility across the boundary between the 'white-collar' classes 1–3 and the other classes, rather than staying within these groups of classes. Next, two variables describe *inheritance* effects, arising from cells where a person is in the same class as his or her father. The first of them is 1 when $r = t = 1$, so it identifies log ORs which involve inheritance of Class 1 (higher managers and professionals). The second captures inheritance effects for all other classes; it is the number of cells where the origin and destination classes are the same but not equal to Class 1; this can be 0, 1 or 2.

Finally, the model includes four *hierarchy* variables which capture effects of different 'distances' of mobility in terms of a hierarchical ordering of the classes. Here Classes 3, 4, and 5 are taken to be on the same level, according to standard practice, so classes $c = \{1, 2, 3, 4, 5, 6, 7\}$ map onto hierarchical positions $h_c = \{1, 2, 3, 3, 3, 4, 5\}$. The variables are defined as $|I_\delta(r, t) - I_\delta(r, u) - I_\delta(s, t) + I_\delta(s, u)|$ where $I_\delta(o, d)$ is an indicator function for a cell defined by origin o and destination d, such that it is 1 if $|h_o - h_d| \geq \delta$, with $\delta = 1, 2, 3, 4$. This is the net count of how many cells in a 2×2 table satisfy the distance condition I_δ, with instances where I_δ holds for both cells of a row or column cancelling out; its possible values are 0, 1 and 2.

The estimated coefficients of the models are shown in Table 5. The results are consistent across cohorts and genders: the strongest explanatory variables are mostly the same in all the models, and each coefficient has the same sign whenever it is firmly determined (with $p < 0.05$; standard errors of the estimated coefficients were evaluated using bootstrap re-sampling of the original individual-level data). The coefficients show that the ratio between indirect and total associations tends to be lower when the total log OR is large. This suggests that there is a limit to how much of the strongest associations can be accounted for by even the largest differences in education between classes of origin. Considering mobility in different parts of the class structure, the ratios are higher for mobility transitions between white-collar and

TABLE 5

Estimated coefficients of linear models for the estimated ratios of indirect to total log odds ratios between origin and destination classes, fitted separately for men and women in each birth cohort. See the text for an explanation of the explanatory variables.

Explanatory variable	1946 cohort		1958 cohort		1970 cohort	
	Men	Women	Men	Women	Men	Women
(Constant)	0.390***	0.630***	0.281***	0.499***	0.309***	0.445***
Total log-OR	-0.208***	-0.194**	-0.232***	-0.308***	-0.295***	-0.225***
White-collar to/from Other	0.147**	0.175*	0.288***	0.236***	0.219***	0.096*
Inheritance:						
Class 1	0.075	-0.093	0.064**	0.014	0.113***	-0.083
Others	-0.110**	-0.122	-0.114***	-0.118**	-0.168***	-0.092*
Hierarchy (distance δ):						
1	-0.049	0.090	0.030**	0.050	0.022*	0.112*
2	0.098***	0.055	0.077***	0.093***	0.089***	0.122***
3	0.155***	0.153*	0.078***	0.128**	0.122***	0.051
4	0.082*	0.038	0.108***	0.251***	0.104***	0.091*
R^2	0.63	0.52	0.58	0.55	0.59	0.33
$m^\dagger$	126	65	260	161	227	122

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Standard errors of the coefficients were calculated using 1000 bootstrap samples of the individual-level data from which the log ORs and their ratios are calculated.
†: The number of ratios used to fit the model. These are the same as in the top part of Table 4.

other classes, for mobility over longer hierarchical distances, and (for men) for comparisons which involve individuals staying in Class 1. The other inheritance effect is negative, so that when a log OR involves one or more cells where a person stays in one of Classes 2–7, the indirect association tends to be smaller than we would otherwise expect.

There are no comparable previous analyses of these data, but two papers have used data from the General Household Survey which represent roughly the same populations, although with different coding of class and education. Kuha and Goldthorpe (2010) used essentially the same method as the one used here, but with the reference distribution $p_0(X = r) = p_0(X = s) = 1/2$ and with just three social classes. They did not examine changes over time, but they too found that the indirect associations were relatively higher for women than for men. Breen and Karlson (2014) considered six classes, using a method based on a latent-variable formulation which estimates different associations from the ones considered here. They considered only men, but compared different birth cohorts; they too concluded that relative strengths of indirect associations had changed little over time.

Our main findings and their sociological implications may be summarised as follows. First, the ratios between indirect and total log ORs that we have estimated predominantly fall below 0.5, indicating that the part that is played by educational attainment in mediating the association between individuals' class of origin and their class of destination is, at most, one of only moderate rather than dominant importance. Other factors that, in total, have at least as great an effect as education must be seen as involved in inequalities of relative mobility chances. Second, the average ratios show no tendency to change across the three birth cohorts that we consider. In other words, there is no evidence that the educational expansion and reform that took place in Britain over this historical period significantly enhanced the role of education in promoting intergenerational class mobility. No movement is apparent towards an education-based meritocracy. Third, the mediating role of education is, however, consistently more important for women than for men, suggesting that in women's working lives more meritocratic processes of social selection are in operation. Fourth, the role of education varies quite markedly in different mobility transitions. In the case of mobility between

Classes 5, 6 and 7 — the largely blue-collar, wage-earning classes at the base of the class structure — the role of education is especially limited; this is also the case with mobility between them and Class 4, small employers and the self-employed, as earlier research has also indicated (Ishida, Müller and Ridge 1995). With longer-range mobility which entails crossing the white-collar/blue-collar divide and other hierarchical levels within the class structure, the mediating role of education becomes of generally greater importance, as might sociologically be expected. But it also emerges that where relative mobility chances are especially unequal, there appears to be a limit to the extent to which such inequalities can be accounted for by the existing differences in educational attainment associated with class origins; other factors, inconsistent with the idea of an education-based meritocracy, clearly supervene.

5. Conclusions. In this paper we have considered the problem of why in Britain the widely held expectation in political and policy circles that educational expansion and reform should increase social mobility has not been realised — these chances having remained essentially unaltered across birth cohorts spanning a quarter of a century. We have found that the extent to which differences in distributions of educational attainment associated with class origins could account for observed class mobility or immobility between generations is by no means so dominant as has been proposed and, most importantly, that while varying with gender and with particular mobility transitions, it shows no tendency to change over time.

Methodologically, this required a method for conducting mediation analysis for population associations between categorical variables. Its ideas and definitions parallel those of causal mediation analysis, essentially just replacing distributions of potential outcomes with conditional distributions of fixed values of variables in a population. The associations thus defined are finite-population parameters which can be interpreted as associations of variables in a suitably standardised population. Estimates of these associations are straightforward, and of a form which could also be used to estimate analogous causal effects in other applications.

These methods could be extended in different ways. For different types of variables, we have considered in Appendix A.3 the case where the mediator M is continuous. If the outcome Y is continuous, the probabilities $p(Y|X,M)$ in Section 3 would be replaced by expected values $\mathrm{E}(Y|X,M)$ from a model for Y (and if the models for M and Y are both linear, our definitions become equal to those of classical linear path analysis). Another extension would be to have multiple mediating variables. If these are treated on an equal footing, so that we wanted to estimate the indirect association via all of them jointly, the conditional distribution of M in the results above is simply replaced with the joint conditional distribution of the mediators. The situation would be more complex if the mediators were treated as being in order, and we wanted to define and estimate associations corresponding to some of the distinct indirect paths that this creates. Indirect associations could in principle be defined analogously, by standardising all conditional distributions which are not on the desired path. Such associations, however, remain to be investigated.

APPENDIX A: ESTIMATION OF THE ASSOCIATIONS AND STANDARD ERRORS

A.1. Matrix expressions of the point estimates. Matrix expressions facilitate the calculation of the log ORs discussed in Section 3 for all values of the variables at once. Recalling that the values of X, M and Y are denoted by $j = 1, \ldots, J$, $k = 1, \ldots, K$ and $l = 1, \ldots, L$ respectively, define the matrix

$$\mathbf{A} = \begin{bmatrix} p(Y=1|X=1,M=1) & \ldots & p(Y=1|X=1,M=K) \\ \vdots & \ddots & \vdots \\ p(Y=L|X=1,M=1) & \ldots & p(Y=L|X=1,M=K) \\ p(Y=1|X=2,M=1) & \ldots & p(Y=1|X=2,M=K) \\ \vdots & \ddots & \vdots \\ p(Y=L|X=J,M=1) & \ldots & p(Y=L|X=J,M=K) \end{bmatrix}, \tag{12}$$

where $p(Y=l|X=j,M=k)$ is in the kth column of the $[(j-1)L+l]$th row, and

$$\mathbf{B} = \begin{bmatrix} p(M=1|X=1) & \ldots & p(M=1|X=J) \\ \vdots & \ddots & \vdots \\ p(M=K|X=1) & \ldots & p(M=K|X=J) \end{bmatrix}. \tag{13}$$

Let $\mathbf{C} = \text{vec}(\mathbf{AB})$ where $\text{vec}(\cdot)$ is the vectorization operator which creates a column vector by stacking the columns of a matrix. Defining the function $c(l,j,j') = (j'-1)JL + (j-1)L + l$, $\mathbf{C}$ is a $J^2L \times 1$ vector whose $c(l,j,j')$th element is $\mathbf{C}[c(l,j,j')] = \sum_{k=1}^{K} p(Y = l|X = j, M = k)\, p(M = k|X = j')$. Let $\mathbf{G}_I$, $\mathbf{G}_D$ and $\mathbf{G}_T$ be $JL \times J^2L$ matrices, defined in such a way that the $[(l-1)J+j]$th row of the matrix (for $j = 1, \ldots, J$, $l = 1, \ldots, L$) has the values $p_0(X = r)$ in its $c(l,r,j)$th columns (for $r = 1, \ldots, J$) and 0 elsewhere for $\mathbf{G}_I$, $p_0(X = r)$ in its $c(l,j,r)$th columns (for $r = 1, \ldots, J$) and 0 elsewhere for $\mathbf{G}_D$, and the value 1 in its $c(l,j,j)$th column and 0 elsewhere for $\mathbf{G}_T$. Let $\mathbf{G} = [\mathbf{G}_I', \mathbf{G}_D' \mathbf{G}_T']'$, and define the $3JL \times 1$ vector $\mathbf{E} = \mathbf{GC} = (\boldsymbol{\pi}^{IE\prime}, \boldsymbol{\pi}^{DE\prime}, \boldsymbol{\pi}')'$. Here $\boldsymbol{\pi}^{IE} = (\pi_1^{IE}(1), \ldots, \pi_1^{IE}(J), \pi_2^{IE}(1), \ldots, \pi_2^{IE}(J), \ldots, \pi_L^{IE}(1), \ldots, \pi_L^{IE}(J))'$ is the vector of all the distinct probabilities $\pi_y^{IE}(x)$, and $\boldsymbol{\pi}^{DE}$ and $\boldsymbol{\pi}$ are similar vectors of all the $\pi_y^{DE}(x)$ and $\pi_y(x)$. Let $\mathbf{F} = \log(\mathbf{E})$, where the logarithm is applied elementwise to $\mathbf{E}$.

The log odds ratios are functions of the elements of $\mathbf{F}$. Consider any $\theta_{rs.tu}^{IE}$ and $\theta_{rs.tu}^{TE}$ where $s > r$ and $u > t$. Let $\mathbf{P}$ be a $2 \times 3JL$ matrix which has values $(1, -1, -1, 1)$ in the columns $\mathbf{v} = ((t-1)J + r, (t-1)J + s, (u-1)J + r, (u-1)J + s)$ of its first row and in the columns $2JL + \mathbf{v}$ of its second row. Then $\mathbf{PF} = (\theta_{rs.tu}^{IE}, \theta_{rs.tu}^{TE})'$ and the ratio $\theta_{rs.tu}^{IE}/\theta_{rs.tu}^{TE}$ can be calculated from these. Finally, point estimates of these quantities are obtained by substituting estimates for the probabilities in $\mathbf{A}$ and $\mathbf{B}$, to obtain estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

A.2. Standard errors of the effects and their ratios. Suppose that the estimated probabilities in $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are asymptotically normally distributed, and that estimates of their asymptotic variance matrices are available. Variance matrices of $\hat{\mathbf{C}} = \text{vec}(\hat{\mathbf{A}}\hat{\mathbf{B}})$ and functions of it are then obtained through repeated application of the multivariate delta method (see e.g. Agresti 2013, Ch. 16). This gives, first,

$$\text{var}(\hat{\mathbf{C}}) = [\mathbf{B}' \otimes \mathbf{I}_{JL}]\, \text{var}[\text{vec}(\hat{\mathbf{A}})]\, [\mathbf{B}' \otimes \mathbf{I}_{JL}]' + [\mathbf{I}_J \otimes \mathbf{A}]\, \text{var}[\text{vec}(\hat{\mathbf{B}})]\, [\mathbf{I}_J \otimes \mathbf{A}]' \tag{14}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{I}_d$ is the $d \times d$ identity matrix (obtaining this requires partial derivatives of $\text{vec}(\hat{\mathbf{A}}\hat{\mathbf{B}})$, see e.g. Lütkepohl 1996). Note that (14) assumes that $\text{cov}[\text{vec}(\hat{\mathbf{A}}), \text{vec}(\hat{\mathbf{B}})] = \mathbf{0}$, which is satisfied here. For $\hat{\mathbf{E}} = \mathbf{G}\hat{\mathbf{C}}$ we then have $\text{var}(\hat{\mathbf{E}}) = \mathbf{G}\text{var}(\hat{\mathbf{C}})\mathbf{G}'$, treating the reference probabilities $p_0(X = x)$ (which are included in $\mathbf{G}$) as fixed quantities. Next, $\text{var}(\hat{\mathbf{F}}) = \hat{\mathbf{S}}^{-1/2}\, \text{var}(\hat{\mathbf{E}})\, \hat{\mathbf{S}}^{-1/2}$, where $\hat{\mathbf{S}} = \text{diag}(\hat{\mathbf{E}})$, and the variance

matrix of the estimate of any $(\theta^{IE}_{rs.tu}, \theta^{TE}_{rs.tu})'$ is $\mathbf{V} = \mathbf{P}\,\mathrm{var}(\hat{\mathbf{F}})\,\mathbf{P}'$ where $\mathbf{P}$ is as defined above, and $\mathrm{var}(\hat{\theta}^{IE}_{rs.tu}/\hat{\theta}^{TE}_{rs.tu}) = \mathbf{d}'\mathbf{V}\mathbf{d}$ where $\mathbf{d} = (1/\hat{\theta}^{TE}_{rs.tu}, -\hat{\theta}^{IE}_{rs.tu}/(\hat{\theta}^{TE}_{rs.tu})^2)'$. Further applications of the delta method can be used if we need covariances of log ORs or their ratios across different values of r, s, t, u. Finally, estimates of these variances and covariances are obtained by plugging in estimates for the conditional probabilities in $\mathbf{A}$ and $\mathbf{B}$.

The specific forms of $\mathrm{var}[\mathrm{vec}(\hat{\mathbf{A}})]$ and $\mathrm{var}[\mathrm{vec}(\hat{\mathbf{B}})]$ in (14) depend on how the conditional probabilities $p(Y|X,M)$ and $p(M|X)$ are estimated. For example, suppose that $p(M=k|X=j) \equiv \gamma_{jk}$ is estimated from a saturated model with $\hat{\gamma}_{jk} = n_{jk}/n_{j\cdot}$ where n_{jk} is the number of sample observations with $(X=j, M=k)$ and $n_{j\cdot} = \sum_k n_{jk}$. Then $\mathrm{vec}(\hat{\mathbf{B}}) = (\hat{\boldsymbol{\gamma}}'_1, \ldots, \hat{\boldsymbol{\gamma}}'_J)'$ where $\hat{\boldsymbol{\gamma}}_j = (\hat{\gamma}_{j1}, \ldots, \hat{\gamma}_{jK})'$ and $\mathrm{var}[\mathrm{vec}(\hat{\mathbf{B}})]$ is block-diagonal with diagonal blocks $\mathrm{var}(\hat{\boldsymbol{\gamma}}_j) = n_{j\cdot}^{-1}[\mathrm{diag}(\hat{\boldsymbol{\gamma}}_j) - \hat{\boldsymbol{\gamma}}_j\hat{\boldsymbol{\gamma}}'_j]$, $j = 1, \ldots, J$. In our analysis in Section 4, where we use a saturated model for both $p(Y|X,M)$ and $p(M|X)$, $\mathrm{var}[\mathrm{vec}(\hat{\mathbf{B}})]$ and $\mathrm{var}[\mathrm{vec}(\hat{\mathbf{A}})]$ are both of this form for the 1958 and 1970 cohorts, whereas for the 1946 cohort they also allow for the fact that the probabilities are estimated using survey weights. If, on the other hand, either of these sets of conditional probabilities are specified using a non-saturated model which depends on some estimated parameters $\hat{\boldsymbol{\phi}}$, the variance matrices are obtained by a further application of the delta method, treating the probabilities as functions of $\hat{\boldsymbol{\phi}}$.

A.3. Estimation when the mediator is continuous. Here we sketch one way of estimating the associations and their standard errors when the mediating variable M is continuous. The discussion is brief because we do not consider this case in our application. In this situation, the sums over M in expressions like (5)–(8) and (10)–(11) are replaced by integrals. In the notation of Appendix A.1, we then have

$$\mathbf{C}[c(l,j,j')] = \int p(Y=l|X=j,M)\,p(M|X=j')\,dM. \tag{15}$$

Everything that uses $\mathbf{C}$ remains unchanged, so that $\pi^{IE}_y(x) = \sum_{x^*} \mathbf{C}[c(y,x^*,x)]\,p_0(X=x^*)$ as in (8), for example, and the indirect log OR is still given by (9), using these $\pi^{IE}_y(x)$.

Models need to be specified for the conditional distributions, and estimates $\hat{p}(M|X)$ and $\hat{p}(Y|M,X)$ are obtained by plugging in estimates of their parameters. To estimate the integrals, we propose to use simple Monte Carlo integration, which approximates (15) by

$$\hat{\mathbf{C}}[c(l,j,j')] = Q^{-1}\sum_{q=1}^{Q} \hat{p}(Y=l|X=j,M_q) \tag{16}$$

for each $j, j' = 1, \ldots, J$ and $l = 1, \ldots, L$, where M_q, $q = 1, \ldots, Q$, are independent random draws from $\hat{p}(M|X=j')$.

The general form of the variance matrix of $\hat{\mathbf{C}}$ defined by (16) is described in Appendix A of Kuha and Goldthorpe (2010), who also give specific formulas for the case where $p(M|X)$ is specified by a linear model and $p(Y|M,X)$ by a multinomial logistic model. This involves uncertainty both from the estimated model parameters and the Monte Carlo integration; the latter can be made small by increasing Q (in Kuha and Goldthorpe 2010, this means omitting all but the last term on the right-hand side of their equation (20)).

More generally, simulation of values from the estimated conditional distributions provides a very flexible way of estimating quantities like these. For example, Imai, Keele and Tingley (2010) propose such an approach for comparable estimation of causal mediation effects (combined with parametric or non-parametric bootstrap estimation of standard errors).

Acknowledgements. This research was supported by funding from the Economic and Social Research Council (grant ES/I038187/1).

SUPPLEMENTARY MATERIAL

Pseudodata, code for data analysis, and results of a variant analysis. The supplement includes a representative pseudo version of the data and R functions and code for its analysis. The values of the estimates from it are similar to the ones for men in the 1970 cohort. Also provided as supplementary materials are results of the analysis of the real data obtained with a different choice of the reference distribution $p_0(X)$, as discussed in Section 4.
().

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, Third ed. Wiley, New York.

BARON, R. M. and KENNY, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

BLALOCK, H. M. (1964). *Causal Inferences in Nonexperimental Research*. The University of North Carolina Press, Chapel Hill, NC.

BLAU, P. M. and DUNCAN, O. D. (1967). *The American Occupational Structure*. Wiley, New York.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

BREEN, R., HOLM, A. and KARLSON, K. B. (2013). Total, Direct, and Indirect Effects in Logit and Probit Models. *Sociological Methods & Research* **42** 164–191.

BREEN, R. and KARLSON, K. B. (2014). Education and Social Mobility: New Analytical Approaches. *European Sociological Review* **30** 107–118.

BREEN, R. and MÜLLER, W., eds. (2020). *Education and Intergenerational Social Mobility in Europe and the United States*. Stanford University Press, Stanford, California.

BUKODI, E. and GOLDTHORPE, J. H. (2016). Educational Attainment – Relative or Absolute — as a Mediator of Intergenerational Class Mobility in Britain. *Research in Social Stratification and Mobility* **43** 5–15.

BUKODI, E., GOLDTHORPE, J. H. and KUHA, J. (2017). The Pattern of Social Fluidity within the British Class Structure: A Topological Model. *Journal of the Royal Statistical Society, Series A* **180** 841–862.

BUKODI, E. and GOLDTHORPE, J. H. (2018). *Social Mobility and Education in Britain: Research, Politics and Policy*. Cambridge University Press, Cambridge.

BUKODI, E. and PASKOV, M. (2020). Intergenerational Class Mobility among Men and Women in Europe: Gender Differences or Gender Similarities? *European Sociological Review*. doi: 10.1093/esr/jcaa001.

BUKODI, E., PASKOV, M. and NOLAN, B. (2020). Intergenerational Class Mobility in Europe: A New Account. *Social Forces* **98** 941–972.

BUKODI, E., GOLDTHORPE, J. H., WALLER, L. and KUHA, J. (2015). The Mobility Problem in Britain: New Findings from the Analysis of Birth Cohort Data. *British Journal of Sociology* **66** 93–117.

BUKODI, E., GOLDTHORPE, J. H., JOSHI, H. and WALLER, L. (2017). Why Have Relative Rates of Class Mobility Become More Equal among Women in Britain? *British Journal of Sociology* **68** 512–532.

DAVIS, J. (1980). Contingency Table Analysis: Proportions and Flow Graphs. *Quality and Quantity* **14** 117–153.

DENIS, D. J. and LEGERSKI, J. (2006). Causal Modeling and the Origins of Path Analysis. *Theory & Science* **7**.

DIDELEZ, V., DAWID, A. P. and GENELETTI, S. (2006). Direct and Indirect Effects of Sequential Treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* 138–146. Association for Uncertainty in Artificial Intelligence Press, Arlington.

DUNCAN, O. D. and HODGE, R. W. (1963). Education and Occupational Mobility: A Regression Analysis. *American Journal of Sociology* **68** 629–644.

ELLIOTT, J. and SHEPHERD, P. (2006). Cohort Profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology* **35** 836–843.

ERIKSON, R. and GOLDTHORPE, J. H. (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon Press, Oxford.

OFFICE FOR NATIONAL STATISTICS (2005). *The National Statistics Socio-economic Classification: User Manual*. Palgrave-Macmillan, Basingstoke.

GENELETTI, S. (2007). Identifying Direct and Indirect Effects in a Non-counterfactual Framework. *Journal of the Royal Statistical Society, Series B* **69** 199–215.

GOLDTHORPE, J. H. (2007). *On Sociology* **2**, Second ed. Stanford University Press, Stanford, CA.

HECKMAN, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* **46** 931–959.

HELLEVIK, O. (1984). *Introduction to Causal Analysis*. George Allen & Unvin, London.

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A General Approach to Causal Mediation Analysis. *Psychological Methods* **15** 309–334.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

ISHIDA, H., MÜLLER, W. and RIDGE, J. (1995). Class Origin, Class Destination and Education: A Cross-National Study of Industrial Relations. *American Journal of Sociology* **101** 145–193.

KUHA, J. and GOLDTHORPE, J. H. (2010). Path Analysis for Discrete Variables: The Role of Education in Social Mobility. *Journal of the Royal Statistical Society A* **173** 351–369.

LOEYS, T., MOERKERKE, B., DE SMET, O., BUYSSE, A., STEEN, J. and VANSTEELAND, S. (2013). Flexible Mediation Analysis in the Presence of Nonlinear Relations: Beyond the Mediation Formula. *Multivariate Behavioral Research* **48** 871–894.

LÜTKEPOHL, H. (1996). *Handbook of Matrices*. Wiley, Chichester.

MITNIK, P., CUMBERWORTH, E. and GRUSKY, D. (2016). Social Mobility in a High-Inequality Regime. *The Annals of the American Academy of Political and Social Science* **663** 140–184.

PEARL, J. (2001). Direct and Indirect Effects. In *Proceedings of the 17th Conference on Uncertainty inArtificial Intelligence* 411–420. Morgan Kaufmann, San Francisco.

POWER, C. and ELLIOTT, J. (2006). Cohort Profile: 1958 British Birth Cohort (National Child Development Study. *International Journal of Epidemiology* **35** 34–41.

ROBINS, J. M. (2003). Semantics of Causal DAG Models and the Identification of Directand Indirect Effects. In *Highly Structured Stochastic Systems* (P. Green, N. Hjort and S. Richardson, eds.) 70–81. Oxford University Press, Oxford.

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3** 143–155.

ROCKHILL, B., NEWMAN, B. and WEINBERG, C. (1998). Use and Misuse of Population Attributable Fractions. *American Journal of Public Health* **88** 15–19.

ROSE, D. and PEVALIN, D., eds. (2003). *A Researcher's Guide to the National Statistics Socio-economic Classification*. Sage, London.

STATACORP (2017). Command **margins**. In *Stata 15 Base Reference Manual* Stata Press, College Station, TX.

TUKEY, J. W. (1954). Causation, Regression and Path Analysis. In *Statistics and Mathematics in Biology* (O. Kempthorne, T. A. Bancroft and J. L. Gowen John W andLush, eds.) 35–66. Hafner, New York.

VANDERWEELE, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.

VANDERWEELE, T. J. and ROBINSON, W. R. (2014). On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables. *Epidemiology* **25** 473–484.

VANDERWEELE, T. J., VANSTEELANDT, S. and ROBINS, J. M. (2014). Effect Decomposition in the Presence of an Exposure-Induced Mediator–Outcome Confounder. *Epidemiology* **25** 300–306.

WACHTER, K. W. (2014). *Essential Demographic Methods*. Harvard University Press, Cambridge, MA.

WADSWORTH, M., KUH, D., RICHARDS, M. and HARDY, R. (2006). Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology* **35** 49–54.

WINSHIP, C. and MARE, R. D. (1983). Structural Equations and Path Analysis for Discrete Data. *American Journal of Sociology* **89** 54–110.

WOLFLE, L. M. (2003). The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography. *Structural Equation Modeling* **10** 1–34.

WRIGHT, S. (1921). Correlation and Causation. *Journal of Agricultural Research* **20** 557–585.

XIE, Y. (1989). Structural Equation Models for Ordinal Variables: An Analysis of Occupational Destination. *Sociological Methods & Research* **17** 325–352.