Unobserved but not unimportant: The effects of unmeasured variables on causal attributions

Robert Coe

School of Education and CEM, Durham University, UK

School of Education, Durham University, Leazes Road, Durham DH1 1TA, UK.
Email: r.j.coe@dur.ac.uk

Objective: To estimate how much difference the inclusion of plausibly important but unmeasured variables could make to estimates of the effects of educational programmes.
Methods: Two examples of policy-relevant research in education were identified. A sensitivity analysis using Monte Carlo simulation was conducted to estimate the size of a possible spurious 'effect' that could actually be entirely due to the failure to incorporate a plausible unobserved variable.
Results: In all the examples the effect size reported in the original study was within the range of possible spurious effects.
Conclusions: What appeared to the original researchers to be substantial and unequivocal causal effects were reduced to tiny and uncertain differences when the effects of plausible unobserved differences were taken into account. Evaluators who rely on statistical control should be more cautious in making causal claims, consider possible effects of unmeasured variables and conduct sensitivity analyses. Alternatively, stronger designs should be used.

Keywords: selection bias, unobserved variables, causal inference, sensitivity analysis, Monte Carlo simulation, education policy

Introduction

Many of those who believe that the most secure basis for causal inference in the social sciences is the evidence from randomised controlled trials would nevertheless concede that causal attributions can sometimes be made on the basis of other methods, and in some cases have to be. The main advantage of random allocation is that all differences between treatment groups, whether observed or not, are controlled. It does not matter whether we have anticipated every possible relevant factor and either explicitly matched the groups on it or measured it adequately and controlled for it statistically; random allocation ensures that the effects of such factors will be equal in all groups (or at least differ only by chance) and will therefore cancel out, within known statistical limits.

By contrast, many analyses of non-randomised designs depend on the assumption that any differences are either fully known or irrelevant. If they are fully known, their effects can be modelled and any residual effects thus attributed to the manipulated variable. If they are irrelevant, by definition, we do not need to worry about them. In practice, the assumption underlying many causal claims from non-randomised designs in education seems to be a combination of the two: differences may not be fully known, but they are known well enough that once we have taken

account of what we do know, any remaining unknown, or inadequately captured, differences can be considered irrelevant.

An exception to this need to either know or ignore unmeasured variables is offered by a class of analytical methods that use instrumental variables. These include methods such as two-stage least squares, multiprocess modelling, structural equation modelling and simultaneous equation modelling (Greene, 2003) which have been widely used in econometrics, and sometimes – though less widely – applied to estimating causal effects in education (eg Steele et al, 2007). These approaches depend on the identification of an 'instrument', *I*, a variable that is correlated with the treatment, *X*, but which has no independent effect on the outcome, *Y*, other than through its effect on the treatment. For this condition to be satisfied, *I* must be uncorrelated with any unobserved factors that influence *Y* (other than through *X*). If such an instrument can be found, the causal effect of *X* on *Y* can be estimated as essentially the ratio of the effect of *I* on *Y* to the effect of *I* on *X* (Winship and Morgan, 1999; Gennetian et al, 2002). However, the assumption that *I* acts only through *X* is generally untestable, sometimes implausible, and even small violations of it can make a big difference to estimates of causal effects, especially if the relationship between *I* and *X* is not strong (Winship and Morgan, 1999, p683; Heckman, 1997; Small and Rosenbaum, 2008; Schneider et al, 2005, p48).

Explicit discussion of the conditions under which causal claims can be justified on the basis of correlational evidence (e.g. Klungel et al., 2004) is perhaps more apparent in health research than in education, though good discussions of a range of approaches relevant to educational research do exist (e.g. Schneider et al, 2007). A review of methods to control for observed and unobserved confounding in non-randomised studies (Groenwold et al., 2009) concludes that unobserved variables cannot be controlled for statistically and implores in its title that 'Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies'.

Other strategies identified by Groenwold et al., and a similar review by Klungel et al. (2004), include the use of matching either in data collection or analysis (eg propensity score analysis) and multivariate analysis. However, all these methods assume that any initial differences between comparison groups are fully known; the possible existence of unmeasured (or inadequately measured) differences poses more of a threat. In discussing the use of instrumental variables approaches, a technique that, as discussed above, potentially deals with the problem of unobserved factors, both reviews conclude that in practice adequate instrumental variables are very difficult to find in health evaluations. Hence their recommendation, as expressed in Groenwold et al.'s (2009) title, is that researchers should conduct sensitivity analyses to quantify the extent to which unmeasured variables could produce effects similar to those observed, under a range of plausible assumptions. As Rosenbaum (2004) explains,

> A sensitivity analysis replaces the statement—'association does not imply causation'—by a specific statement about the magnitude of hidden bias that would need to be present to explain the associations actually observed. (p10812)

The stimulus for the current investigation was research in the field of education that makes causal, policy-relevant claims on the basis of non-randomised designs. In educational research these designs are widely used in impact evaluation studies and their interpretation seems often to be treated as unproblematic. Indeed, there is considerable opposition in principle from some quarters to the use of randomised trials (e.g. Morrison, 2009, Goldstein, 1987). An interesting comparison can be made with the field of health research, where the randomised controlled trial is

widely accepted as providing the 'gold standard' of evidence of causal effects (eg Klungel et al., 2004; Rubin, 2008), although non-randomised designs are also widely used in epidemiological studies.

Numerous examples of the use of sensitivity analyses to estimate the effects of unmeasured variables can be found in health research, but in educational research they are much harder to discover. An early paper by Rosenbaum (1986) applied the technique to a comparison of the achievement of matched pairs of high school drop-outs and controls. More recently, Leow et al. (2004) used this approach to help interpret a comparison on basic skills performance of those who had and had not taken advanced courses.

Although formulae are available to estimate the size of spurious effects due to unmeasured variables, all seem to have limitations for this context. Some (eg Lin et al., 1998; Arah et al., 2008) are restricted to the case where the outcome variable is dichotomous, as is common in health research. Others (eg Rosenbaum, 1991) require the combined unmeasured variables to be effectively dichotomous, or depend on the analysis of matched pairs (Rosenbaum, 1986). Still others (Pan and Frank, 2003) test dichotomous decisions of whether causal effects are significant or not and require complex calculations to implement.

Instead, the current study uses Monte Carlo simulation to estimate the possible effects of plausible unmeasured variables on estimates of the causal impact of an educational programme derived from a non-experimental analysis. Two examples of policy evaluation in education have been chosen to illustrate the method. These studies were not the result of considering a large number of studies and selecting a small number to make a particular point: on the contrary, the first examples tested proved to illustrate the point nicely. A third example, an evaluation of the effects of the Assisted Places Scheme (means-tested payment of fees at private schools; Powers et al., 2006) was also used, but has been omitted for reasons of space (see Coe, 2009 for an extended report). All the examples were chosen initially because they have been used to make relatively unequivocal, policy-relevant, causal claims that appeared to me on reading them to be rather more confident than seemed justified. The fact that I was unable to translate my feeling of unease about their confident assertions into a clear critique was the motivation for developing and applying the simulation.

Design of the simulation

Model simulated

The model simulates a situation where the effect of a binary group membership or 'treatment' variable, X, on an outcome measure, Y, is estimated with a set of measured covariates, M, using OLS linear regression:

$$Y = \alpha_0 M + \alpha_1 X + e_1$$

Equation 1

where α_0 and α_1 are regression coefficients and e is the residual error. For simplicity, M is taken as a single variable, though this could be thought of as a linear sum of a set of covariates. In this case instead of the correlation between M and Y we could talk about the multiple correlation.

However, the true value of Y is given by:

$$Y = \beta_{} M + \beta_{} U + e_2$$

Equation 2

where U is an unmeasured covariate that is associated with both group membership and the outcome.

In other words, the outcome is actually determined (within random error, e_2) by a combination of measured and unmeasured variables. Hence the coefficient of group membership (α_1 in Equation 1) represents a spurious group effect, an artefact of the association between X and U and the failure to include U in the model. The residual, e, is assumed to be $N(0,\sigma^2)$.

Parameters for the simulation

We must allow for all possible inter-correlations among the variables, Y, M and U. In other words,

p_{uy}	Correlation between Outcome, Y, and Unmeasured covariate(s), U.
q_{um}	Correlation between Unmeasured covariate(s), U, and Measured covariate(s), M.
r_{ym}	Correlation between Outcome, Y, and Measured covariate(s), M.

We must also allow the strength of relationship between U and X to vary:

s_{ux}	Correlation (point-biserial) between Unmeasured covariate(s), U, and (binary) Group membership, X.

And finally, we want to know what values of the apparent but spurious group effect are possible for each combination of these:

E	Phantom effect of Group membership, X, on Outcome, Y (ie α_1 above).

In practice, we are likely to know E and r_{ym} since these have been estimated in the (under-specified) regression model that has been fitted (Equation 1). We need to identify what possible, plausible or likely values of p_{uy}, q_{um} and s_{ux} could produce these known values even if there is no true effect of X on Y.

Setting up the simulation

The simulation was run in SPSS 15.0. Initially, 100,000 cases of five normally distributed N(0,1) random variables (RV1 to RV5) were generated. These were used to compute the variables Y, U, M and X with the desired inter-correlations. The equations used and the SPSS syntax can be found in the Appendix. Values of p_{uy} and q_{um} were selected to match plausible estimates of those in the example studies. The value of s_{ux} was allowed to vary by adding a random variable to U with different (random) relative weights and allocating $X = 1$ if the sum was greater than zero, otherwise $X = 0$. For each assigned combination of p_{uy}, q_{um} and r_{ym}, the simulation was run 50 times, each with a different (random) value of s_{ux}. For each of these simulations, the value of E, the spurious phantom 'effect' that would appear in a

regression model that omitted U, was plotted against s_{ux} and a trend line (cubic polynomial) fitted (see Figure 1 to Figure 3).

At one extreme (s_{ux} =0), group membership (X) was purely at random, uncorrelated with the unmeasured variable, U. At the other extreme, group membership was entirely determined by the value of U: $X = 1$ if $U > 0$, $X = 0$ otherwise. Note that even in the latter case of complete dependence, if a Gaussian variable is dichotomised like this the correlation (s_{ux}) between U and X is capped at about 0.8. In between these two extremes are varying levels of dependence between the unmeasured variable and group membership.

One of the most difficult aspects of interpreting the results of the simulations is to estimate a plausible value for this correlation, s_{ux}. Table 1 shows a conversion between values of this correlation, the equivalent standardised effect size and an interpretation based on Rosenthal and Rubin's (1982) binomial effect size display (BESD). This illustrates how if the population were dichotomised at the mean value on the unmeasured variable, U, the possible percentages of high and low scorers who would be found in each treatment group. Various assumptions are required for this conversion, which is inevitably simplistic, but may nevertheless provide some help in interpretation (for further explanation and discussion see Coe, 2002).

Table 1: Illustrative interpretations of different values of the correlation, s_{ux}, between group membership, X, and the unmeasured variable, U.

Correlation (point-biserial) between group membership, X, and unmeasured variable, U.	Equivalent standardised effect size	Percentage of those above average on the unmeasured variable, U, who are in	
		the intervention group	the control group
s_{ux}	d	$X = 1$	$X = 0$
0	0.00	50%	50%
0.1	0.20	55%	45%
0.2	0.41	60%	40%
0.3	0.63	65%	35%
0.4	0.87	70%	30%
0.5	1.15	75%	25%
0.6	1.50	80%	20%
0.7	1.96	85%	15%
0.8	2.67	90%	10%

Summary of the example studies

1. The Impact of Study Support

Background to the study

This study by MacBeath et al (2001) is subtitled 'A report of a longitudinal study into the impact of participation in out-of-school-hours learning on the academic attainment, attitudes and school attendance of secondary school students'. The report analysed the performance at GCSE and Key Stage 3 (national assessments at age 16 and 14) of over 8000 pupils in 52 schools, as well as substantial qualitative data. The main statistical analysis used multiple regression. The report was commissioned and published by the Government Department for Education and Skills. It is still (September 2009) available and cited as key support for a number of related policy initiatives (see http://www.standards.dcsf.gov.uk/studysupport/about/).

Analysis and results

The study evaluated a variety of forms of study support including Y10 subject-focussed, Y10 sport, Y10 aesthetic, Y11 subject-focussed, Y11 other, Y10 drop-in and Y11 Easter school. Of these the last had the biggest effect on achievement and is therefore taken as an upper bound for the effect. Outcomes included GCSE English, GCSE mathematics, average of best 5 GCSE grades, number of A-C passes, Key Stage 3 average, as well as attitudes and attendance. The first two of these have been chosen for this simulation as they are simple, valued outcomes that do not require the assumption that all GCSEs are equally difficult and span the range of R^2 values quoted (63.1% and 70.4% respectively). Covariates in the model are Y9 SATs (national tests in English, mathematics and science), gender and individual school dummy variables. Regression coefficients cited in the study for the 'effects' of Y11 Easter School are equivalent to standardised mean differences of 0.18 for English and 0.11 for mathematics.

Claims made by the study

The word 'impact' in the title announces the causal claim unequivocally. Almost every sentence in the executive summary (p7) confirms this, for example,

> 'Study support has effects which are significant and substantial for GCSE performance'
> 'Study support can improve attainment in Maths and English by half a grade'
> 'Participation improves Maths attainment by one third of a level and Science attainment by three quarters of a level'
> 'All students who participate benefit from study support'
> 'Participation in study support has a favourable effect on attitudes to school'
> 'participation in some forms of study support has a positive impact on school attendance'
> 'Study support has an impact at whole school level when participation rates are high'

2. Gifted and Talented support in 'Excellence in Cities'

Background to the study

The *Excellence in Cities* (EiC) initiative was launched in 1999 with the aim of raising standards in inner cities and other urban areas in England (Kendall et al, 2005; see

also DCSF, 2009). There were seven key policy strands, of which 'Gifted and Talented' is our focus here. This strand became national policy before the end of EiC and continues to be a key element of government policy for schools in England (see http://www.standards.dcsf.gov.uk/giftedandtalented/). As part of the Gifted and Talented (G&T) strand, schools were asked to identify between 5 and 10% of their most able pupils. Schools were given additional funding to support the learning of these pupils, most of which was spent on specialist teaching materials, teacher salaries or incentive points, out-of-school activities and supply cover (Kendall et al., 2005, p91).

Analysis and results

Detailed analysis of the attainments of pupils are presented in a report by Morris and Rutt (2005). Attainment outcomes included Key Stage 3 levels in mathematics and English, as well as average level, total GCSE score, capped 8 (total of the best 8 GCSE grades) and average GCSE grade. The highest R^2 values in multilevel models at each Key Stage were for average KS3 level (66%) and capped 8 GCSEs (80%) (p20). Regression coefficients (fixed effects) were estimated for 42 different explanatory variables, including a dummy for G&T designation. Conversion of coefficients of G&T gives standardised effect sizes of 0.30 for average KS3 and 0.22 for capped 8 GCSE.

Claims made by the study

Although some caveats and limitations of the data are pointed out by Kendall et al (2005), in the main summaries of the results the causal attribution is clear. For example:

> "… EiC has led to an increase in average attainment" (p25)
> "… there was evidence of a positive impact for some specific groups of students" (p25)
> "This was equivalent to increasing the percentage of pupils achieving level 5 or above by between 1.1 and 1.9 percentage points" (p25)
> "…early mentoring (in Year 7) had enabled some pupils to overcome barriers to learning" (p26)
> "The impact of the [G&T] Strand appeared greater for pupils with lower levels of attainment at the end of Key Stage 3." (p54)
> "… the main effects from the quantitative analysis have been in three areas:
> • in improving levels of attainment in Mathematics at Key Stage 3 (with the greatest impact in the most disadvantaged schools)
> • for pupils identified as gifted and talented
> • in improving attendance." (p121)

Results of the simulations

1. The Impact of Study Support

Assigning values of the inter-correlations p, q, r

The R^2 values quoted above for this study (MacBeath et al, 2001) correspond to multiple-R values of 0.79 for English and 0.84 for mathematics. Hence an appropriate value for r_{ym}, the correlation between the outcome (GCSE grade) and measured

covariates (KS3 SATs, gender, and school level dummies) is 0.8. This represents a high level of explained variance for an outcome measure such as GCSE performance.

Identifying possible unmeasured variables that might account for the effect is to some extent a matter of speculation. One obvious candidate would be socioeconomic status (SES). A meta-analysis by Sirin (2005) found the average correlation between SES and academic achievement to be around 0.3. However, in practice SES measures are often a rather loose and poorly measured proxy for the true family situation; the use of SES measures that are unrestricted, focus on home resources (eg books), collect data directly from parents and relate to specific outcomes such as mathematics or verbal measures can increase the correlation to as much as 0.5. It certainly seems plausible that SES could influence students' likelihood of participating in study support activities such as Y11 Easter school: students from more advantaged backgrounds could well be more likely to take part.

Other possible candidates for unmeasured variables in this study include motivation (Ugoroglu and Walberg, 1979, found average correlation with achievement of 0.34), self-discipline (found by Duckworth and Seligman, 2005, to have correlations as high as 0.7 with achievement), emotional intelligence (Goleman, 1998), resilience (Wang et al, 1998), self-esteem (Brookover et al, 1965), educational aspirations (Sewell and Shah, 1968). All these have been shown to be related to academic achievement and could plausibly influence the chances of a student embarking on and sustaining participation in out-of-lesson study support. There is no need for us to consider school-level factors, as the inclusion of school dummy variables will have dealt with any unobserved school-level confounders. However, it is possible that within-school variations in teacher-level characteristics could confound the estimate of effects. For example, the level of individual teachers' motivation, commitment and general instructional quality could influence both GCSE outcomes and their students' willingness to attend an Easter school.

Taken individually, several of these factors have been shown to have correlations of at least 0.3 with achievement. Of course, there may be considerable overlap among them, but if all these factors were measured well and combined, it is likely that the overall correlation with achievement would exceed 0.3. Certainly, a multiple correlation of around 0.4 does not seem hard to accept and 0.5 or higher may even be defensible.

Given these considerations, the simulation was run with estimates of the correlation between unmeasured variables (U) and both prior and outcome achievement (M and Y, respectively) of 0.3, 0.4 and 0.5.

Results

The results of the simulations are shown in Figure 1. If the unmeasured variables in the model are assumed to have a correlation with both outcome and prior attainment of around 0.3, the artefactual attribution of the true effects of this variable to group membership ranges from zero to an effect size of about 0.10, depending on the strength of the relationship between group membership and the unmeasured variable. If the correlation is 0.4 it ranges from zero to 0.14 and climbs to 0.19 if the correlation is assumed to be 0.5.

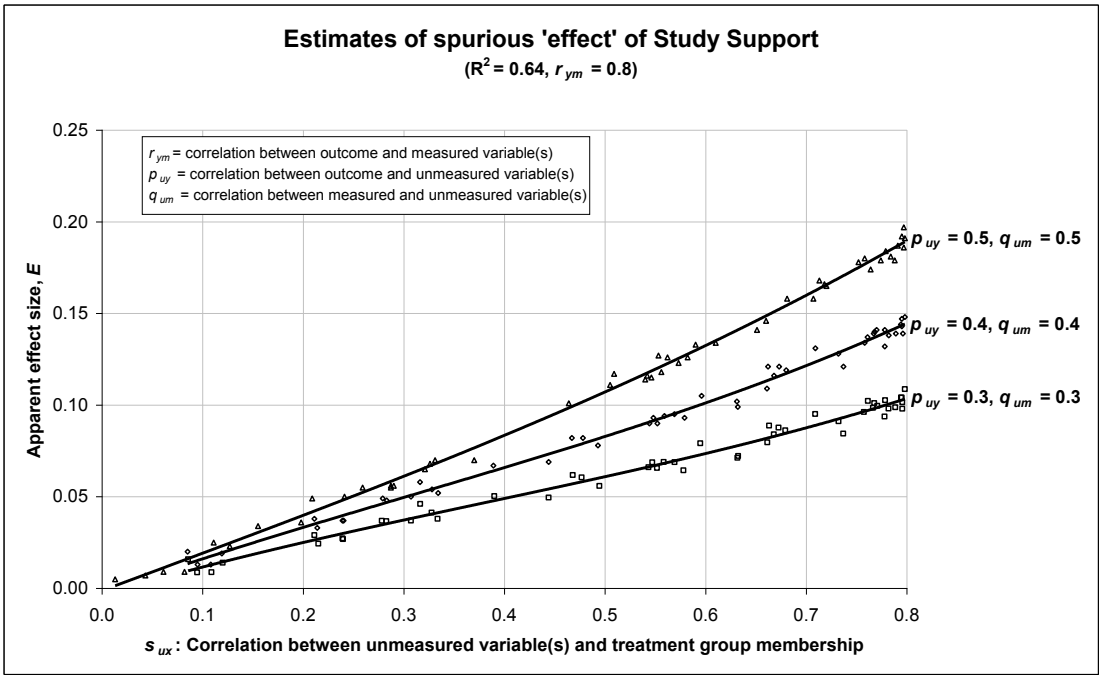

Figure 1: Relationship between s_{ux}, the strength of association between unmeasured variable and group membership, and the spurious 'effect' , E, that would be attributed to Study Support participation, under different assumed values of p_{uy} and q_{um} (the correlations between the unmeasured variable with outcome and measured covariates respectively), when r_{ym} (the correlation between measured covariates and outcome) is fixed at 0.8.

Interpretation

In the MacBeath et al (2001) study, the effect sizes for 'Y11 Easter school' on GCSE English language and on GCSE mathematics were estimated at 0.18 and 0.11 respectively (see above).

The former value (0.18) seems to be right at the top end of what could be an artefact of inadequate modelling, under the plausible assumptions outlined above. Failure to include variables such as socioeconomic status, motivation or teacher quality in the model might have inflated the estimate of the effect of study support in this case, but unless we are willing to assume the maximum plausible correlation between these unmeasured variables and GCSE English grade, and to posit an extremely strong relationship between the unmeasured variables and participation in the Y11 study school, this failure cannot account for the whole of the effect.

We must try to narrow the range of plausible assumed parameters. Estimating plausible values for s_{ux}, depends on knowing the likely strength of the relationship between unmeasured variables, such as SES, and participation in Y11 study school. Data from the Yellis (Year 11 information system) survey (www.yellisproject.org) suggest correlations of the order of 0.3 between measures of either socioeconomic status or academic motivation with reported participation in after-school clubs. It seems likely that for an activity such as the Easter revision course, the association between participation and academic motivation could be higher than this. Hence we may assume that s_{ux}, the correlation between the unmeasured factors and participation in study support, may be in the range 0.2 to 0.5. For the values of p_{uy} and q_{um} , any of the values used in the simulation (0.3, 0.4 and 0.5) seem plausible. With this range of likely parameters, the range of possible values for *E*, the estimate of the spurious effect is from 0.03 to 0.11.

If we go further and seek a single 'best guess' for E we might take 0.4 as the most likely value for p_{uy} and q_{um} and 0.4 as an estimate for s_{ux}. This would produce a spurious effect size of about 0.07. In this case, the true effect of Y11 Easter school study support on GCSE English would be the difference between this and the estimate (0.18) from the regression model, i.e. about 0.11.

For GCSE mathematics, on the other hand, the effect size of 0.11 appears to be just within the range that might plausibly occur as an artefact of underspecifying the model. Under the same 'best guess' that led to estimating the artefactual effect as 0.07, the true effect would remain at only 0.04. Although positive, this is a very small effect, unlikely to be statistically or educationally significant and, given the uncertainty surrounding many of the assumptions made, inevitably subject to a wide margin of error.

2. Gifted and Talented support in 'Excellence in Cities'

Assigning values of the inter-correlations p, q, r

Multiple-R values corresponding to the R^2 values cited above are 0.8 for average KS3 and 0.9 for capped 8 GCSE. These two values were therefore used for the correlation between outcome and measured variable, r_{ym}.

Given such a high proportion of variance explained in the model, and such a wide range of explanatory variables included, it might seem unlikely that any unmeasured variable could make much difference to the outcome. However, the particular nature of the G&T designation opens up another type of threat here. If pupils are identified as G&T on the basis of their attainment, then they are effectively selected for 'treatment' on a variable that is highly correlated with the outcome. For example, suppose participation in Year 11 G&T activities ($X = 1$) was open to those who had performed in the top 10% in a school on their internal end of Y10 exams. If we take these Y10 exam results as the unmeasured variable, U, we would expect very high correlations between U and X. Plausible values for the correlation between the measured covariates (prior attainment, FSM status, gender, ethnicity) and unmeasured (Y10 exam) are likely to be 0.7 or 0.8, as are correlations between the latter and the outcome measure (GCSE). Similar correlations could be expected for KS3, where the unmeasured variable would be Y8 exam performance, though as the multiple-R value for KS3 is slightly lower, we may reduce the other correlations accordingly.

Hence for KS3, the simulation was run with $r_{ym} = 0.8$; $p_{uy} = 0.6$ and 0.7; $q_{um} = 0.6$ and 0.7. For GCSE, the simulation was run with $r_{ym} = 0.9$; $p_{uy} = 0.7$ and 0.8; $q_{um} = 0.7$ and 0.8.

Results

The results of the simulations for the impact of G&T on KS3 are shown in Figure 2 and on GCSE in Figure 3.

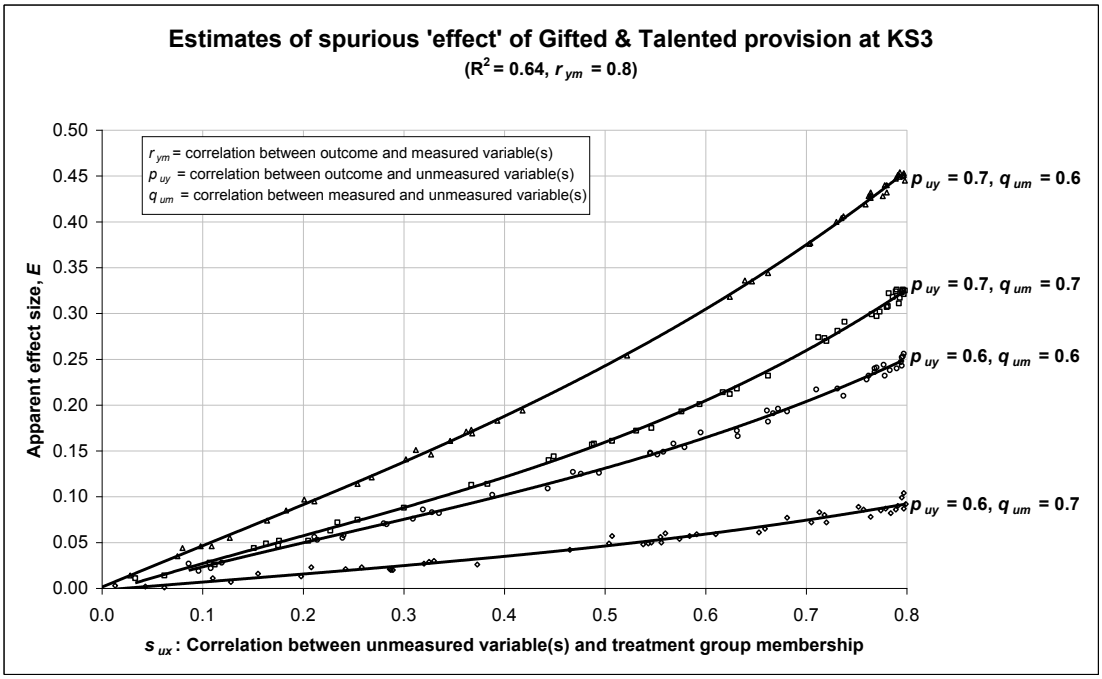

Figure 2: Relationship between s_{ux}, the strength of association between unmeasured variable and group membership, and the spurious 'effect' , E, that would be attributed to G&T participation, under different assumed values of p_{uy} and q_{um} (the correlations between the unmeasured variable with outcome and measured covariates respectively), when r_{ym} (the correlation between measured covariates and outcome) is fixed at 0.8.

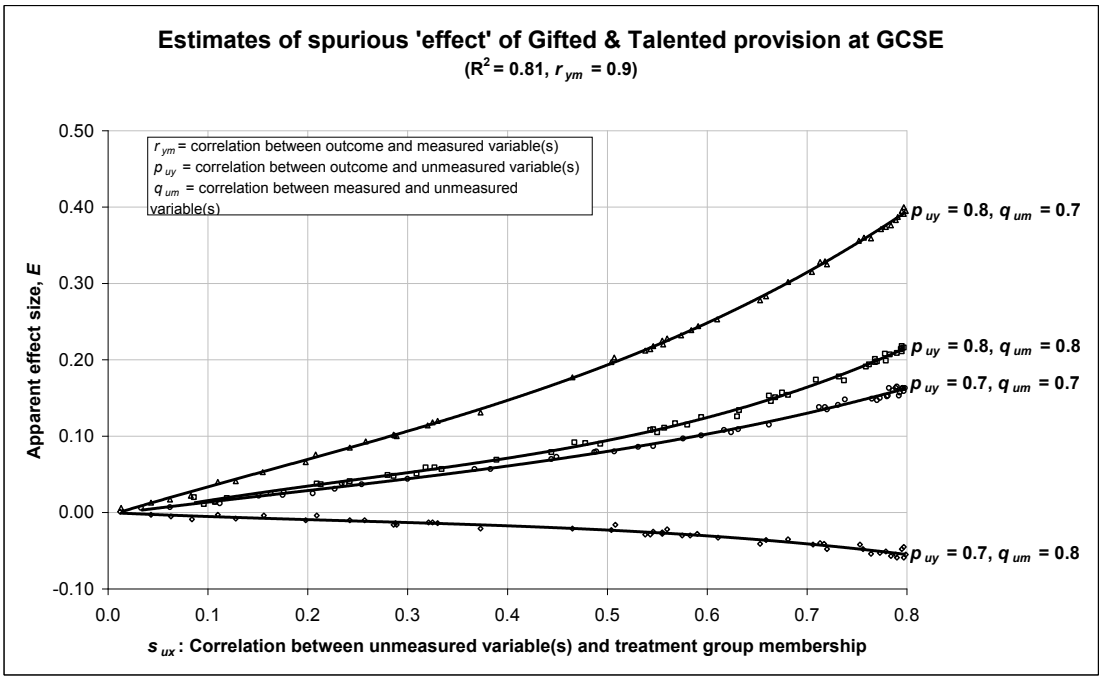

Figure 3: Relationship between s_{ux}, the strength of association between unmeasured variable and group membership, and the spurious 'effect' , E, that would be attributed to G&T participation, under different assumed values of p_{uy} and q_{um} (the correlations between the unmeasured variable with outcome and measured covariates respectively), when r_{ym} (the correlation between measured covariates and outcome) is fixed at 0.9.

Interpretation

For the outcome average KS3 level (with $r_{ym} = 0.8$), it is clear that even with relatively high proportions of variance explained by the model, if an unmeasured variable has high correlation (p_{uy}) with the outcome, then quite substantial spurious effects can appear. For $p_{uy} = 0.7$ the artefact could be equivalent to a standardised effect size as high as 0.45; up to 0.25 if $p_{uy} = 0.6$. Reducing the estimate of the correlation (q_{um}) between the hypothesised unmeasured variable and the measured variables already in the model also increases the size of the spurious effect. This might correspond to changing the timing of the assessment that is used to determine G&T status: if it is close to the baseline assessment, q_{um} would be high, p_{uy} would be lower; if it is closer to the outcome assessment, q_{um} would be lower, while p_{uy} would be high.

For this simulation it seems plausible that the correlation (s_{ux}) between G&T status (X) and the unmeasured Y8 assessment (U) could be very high, since eligibility for G&T might conceivably be entirely dependent on performance (ie $s_{ux} = 0.8$). Even if other factors are taken into account in identifying G&T pupils, s_{ux} seems unlikely to fall below 0.6. Under these assumptions, the actual effect size estimated by the multilevel model (0.30) is just about within the range that could be a pure artefact of selecting treatment groups on the basis of an unmeasured variable that is highly correlated with the outcome: E is between 0.16 and 0.32. If we had to make a best guess at a single combination of parameters, we might choose $p_{uy} = 0.7$, $q_{um} = 0.7$, $s_{ux} = 0.7$, giving an estimate of the spurious effect of $E = 0.26$. The difference between this and the actual estimate is too small to be considered evidence of a genuine effect.

For effects on 'GCSE capped 8', the proportion of variance accounted for in the model is even higher ($r_{ym} = 0.9$). Even so, if we are prepared to hypothesise that a pupil's designation as G&T could be strongly dependent on an assessment whose results were not available to the evaluators, but that was itself highly correlated with the outcome, then we could still get a substantial spurious effect: at the top end of our plausible assumptions, a standardised effect size of 0.40 is possible. Here a 'best guess' set of parameters might be $p_{uy} = 0.8$, $q_{um} = 0.8$, $s_{ux} = 0.7$, giving an estimate of the spurious effect of $E = 0.17$. Again, this is below the effect size actually estimated (0.22) so there may be a remaining positive effect, but the difference (0.05) is certainly small and subject to a good deal of uncertainty.

Interestingly in this model, if the correlation (p_{uy}) between the outcome and this unmeasured variable falls by just 0.1, other things being equal, we will see a negative spurious effect emerging. Under these conditions the estimated effect size might actually be smaller than the true effect. Hence it seems the size of the spurious effect is quite sensitive to small changes in the parameters assumed.

Discussion

Two examples have been used to illustrate that when regression models are used to estimate the effects of participation in a programme by adjusting for known covariates, these estimates can, under reasonable assumptions, be substantially biased by the failure to include in the model other factors that may be related to both group selection and the outcome. In all the cases considered here, what appeared to the researchers to be a substantial and unequivocal difference interpretable as a causal effect either disappears or becomes reduced to a tiny and uncertain difference when

the effect of unobserved differences is taken into account. A summary of the example studies, their results and claims, together with the results of the simulation and its interpretation is presented in Table 2. For these examples it is clear that the conclusions of the studies would have been very different had the effects of unobserved differences been considered. A number of further points emerge as worthy of comment.

The first is that even statistical models with a very high proportion of variance explained (R^2) can be subject to these spurious effects. One might have thought that R^2 values as high as 0.81 (multiple-R of 0.9) would pretty much guarantee that, with such a small amount of remaining unexplained variance to account for, other factors could not make too much difference. This thought proved to be wrong, however.

Second, the size of these spurious effects is quite sensitive to small changes in the assumed values of the parameters in the simulation. For example, we saw that the estimate of the effect of G&T provision on GCSE could fall from 0.22 down to -0.06 just by changing the estimate of the correlation (p_{uy}) between the outcome and the unmeasured variable by as little as 0.1, other things being equal (see Figure 3).

Third, and related to the second, is that choosing reasonable values for the assumed parameters is far from easy. As a result, the 'best guess', 'likely range' and even 'possible range' of values calculated here for the spurious effect, *E*, will be controversial and open to challenge. No doubt some readers will already have thought as they read the account of the simulations that the assumptions underpinning them were wrong, or at least open to argument.

Fourth, the fact that it is both hard and crucial to get the assumptions right is not a reason not to try. Debates are necessary about issues such as what assumptions are reasonable, what kinds of unmeasured variables should be considered, what ranges of their possible correlations with the measured and outcome variables are plausible, and what the strength of the relationship might be between these unobserved variables and group selection. If there are differences of opinion on these matters then we will be able to see what effect those differences might have on the conclusions from the study. Unless we believe that getting a single, simple answer is more important than getting it right, making this uncertainly explicit is a good thing. We must also bear in mind that not to engage in a debate about these assumptions is in effect to make a default assumption that all the correlations are zero.

Fifth, it is important to point out that the example studies considered here are of generally high quality and excellent in many ways. These are by no means taken from the bottom end of the quality spectrum of evaluations of educational policy initiatives. Certainly, with regard to technical issues such as sampling, instrumentation, survey execution, complementary mixing of quantitative and qualitative data and the sophistication of the methods of statistical analysis used, these studies can claim to be exemplary in at least some respects. Where they are all open to criticism, however, is in their failure to consider other possible causes of the effects they describe.

Table 2

Study	MacBeath et al. (2001)		Kendall et al. (2005)	
Intervention / programme	Study support (Y11 Easter School)		Gifted & Talented provision	
Outcome(s)	GCSE English;	GCSE maths	KS3 average level;	GCSE capped 8 score
Covariates	KS3 SATs average, Gender, School type		Prior attainment, FSM status, gender, ethnicity	
R^2 in the model	63%;	70%	66%;	80%
Estimate of the effect, from regression model	0.18;	0.11	0.30;	0.22
Interpretation given by the researchers	'Study support can improve attainment in Maths and English by half a grade'		'Pupils designated as gifted and talented had higher levels of attainment at the end of Key Stages 3 and 4 than those of otherwise similar pupils not designated.'	
Possible relevant unmeasured variable(s)	Socioeconomic status; Motivation; Self-discipline		Attainment used to identify G&T status	
Range of possible spurious effects	0.0 – 0.19		0.0 – 0.45	-0.06 – 0.40
Range of likely spurious effects	0.04 – 0.11		0.16 – 0.32	0.10 – 0.21
Best guess at spurious effect	0.07		0.26	0.17
Justified conclusion, taking account of bias due to omitted factors	Possible small residual effect (0.11) on English but pretty much no genuine effect on maths		Any genuine effect for both outcomes is very close to zero	

Recommendations for research

A number of recommendations for further work emerge from this study. The first is that there are probably other ways such a sensitivity analysis could have been conducted. Validating the results of this simulation against other methods of achieving the same would be a useful step. If the results prove to be robust, ways of making this kind of approach easier to conduct with widely available software should be explored.

Second, it would not be too difficult to extend the method to other kinds of statistical analyses. In fact, Kendall et al.'s (2005) evaluation used multilevel models rather than the simple OLS regression used in the current simulation; it is possible that a simulation using multilevel models would have arrived at a different result. Propensity score matching has become the method of choice for creating well-

matched groups in non-randomised comparisons in health research and other areas. This approach offers significant advantages over the kinds of multivariate regression used in the educational examples considered here (Klungel et al., 2004). It would certainly be valuable, though perhaps more difficult, to conduct sensitivity analyses of the results of studies that have used this approach.

Thirdly, if the two studies considered here are at all typical, there is a need for educational researchers to be considerably more cautious in making causal claims on the basis of statistical analyses of the differences between those who have and have not experienced a particular educational programme. It seems that the issues that were debated, the lessons learned and the developments made in epidemiology fifty years ago in debating whether non-experimental evidence could establish smoking as a cause of lung cancer (e.g. Cornfield et al, 1959; see Hill et al, 2003) have passed us by: we need to catch up. Educational researchers must also be aware of a range of powerful statistical methods for dealing with selection bias due to unobserved variables developed and widely used by econometricians (Green, 2003). Policy makers who use their results may also need to be more cautious and critical of researchers' claims – assuming they genuinely want to know whether the policy will work (Pritchett, 2002).

Fourthly, evaluators who use statistical control to evaluate effects of educational programmes should give more explicit thought to the possible effects of unmeasured variables. At the very least, to acknowledge the possibility that something other than the programme might be responsible for the difference would be a start. Better still would be to systematically list possible factors that are either unmeasured or inadequately measured and that might be related both to participation in the programme and to the outcome. For each factor, an argument should be made about the plausibility of its influence, based on existing evidence about associations and theoretical arguments about possible mechanisms. Even better, this argument should include a sensitivity analysis to quantify how big an effect it could plausibly have, under particular, plausible and explicitly stated assumptions. As has been demonstrated here, conducting such a sensitivity analysis is no more difficult than the kinds of statistical techniques routinely used by evaluators.

Finally, the extent to which unmeasured factors can undermine causal claims reinforces the case for the use of stronger evaluation designs and analyses. Even if we acknowledge that randomised controlled trials (RCTs) are not always appropriate or possible, it is almost certainly the case that they could be used more often than they are (Torgerson and Torgerson, 2007; Cook, 2003; Slavin, 2008). Furthermore, there are non-randomised designs that are considerably stronger than the ones used in the examples here: for example, regression-discontinuity and time-series designs (Shadish et al., 2002; Cook et al., 2008).

Of course, RCTs can be ethically problematic; they can impose restrictions on the representativeness of samples, interventions or contexts; they may have inappropriate time-frames; and they do not necessarily solve problems such as attrition, implementation fidelity, contamination of treatments or poor outcome measurement. However, their claim to offer the 'gold standard' of evidence for causal inference rests on their power to minimise selection bias. If the threat of this bias is not regarded as a particular problem, as seems to have been the case in the examples considered here, then the arguments against RCTs may seem convincing. The current study suggests, though, that the bias arising from unmeasured factors can be a very significant problem: the interpretation of results and attribution of causality can be

entirely overturned. If so, the case for using a random allocation design (RCT) that can eliminate this bias is more compelling.

References

Arah OA, Chiba Y and Greenland S (2008) Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. ANNALS OF EPIDEMIOLOGY, Volume: 18, Issue: 8, Pages: 637-646.

Brookover, W.B., LePere, J.M., Hamachek, D.E., Thomas, S., Erickson, E.L., 1965, Self-concept of ability and school achievement II. Co-operative research project no. 1636, Michigan State University

Coe, R. (2002) *It's the effect size, stupid: what effect size is and why it is important.* Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002.

Coe, R. (2009) 'Unobserved but not unimportant: The effects of unmeasured variables on causal attributions'. Paper presented at 4th Annual Conference on Randomised Controlled trials in the Social Sciences, University of York, 14-15 Sept 2009.

Cook, T.D. (2003) 'Why have educators chosen not to do randomized experiments?' *Annals of the American Academy of Political and Social Sciences*, 589, 114-149.

Cook, T. D., Shadish,W. R., andWong, V. C. (2008), "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings From Within-Study Comparisons," *Journal of Policy Analysis and Management*, 27, 724–750.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of The National Cancer Institute* Volume: 22 Issue: 1 Pages: 173-203.

DCSF (Department for Children, Schools and Families) (2009) The Standards Site – Excellence in Cities. http://www.standards.dfes.gov.uk/local/excellence/whatis_eic.html (Accessed 12.7.09)

Duckworth, A. L. and Seligman, M. E. P. (2005) Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychological Science*; Dec2005, Vol. 16 Issue 12, p939-944.

Gennetian, L.A., Bos, J.M. and Morris, P.A. (2002) Using Instrumental Variables Analysis to Learn More from Social Policy Experiments. Manpower Demonstration Research Corporation.

Goldstein, H. (1987). Multi-level models in Educational and Social Research.: London : Griffin

Greene, W.H., (2003) *Econometric Analysis*, Fifth edition. New Jersey: Prentice Hall.

Groenwold, R.H.H., Hak, E., and Hoes, A.W. (2009) Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of Clinical Epidemiology*, 62, 22-28.

Heckman, J.J. (1997) Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 32, 3, 441–62.

Hill, G., Millar, W. and Connelly, J. (2003) "The Great Debate": Smoking, Lung Cancer, and Cancer Epidemiology. *Canadian Bulletin of Medical History*, 20, 2, 367-386.

Kendall, L., O'Donnell, L., Golden, S., Ridley, K., Machin, S., Rutt, S., McNally, S., Schagen, I., Meghir, C., Stoney, S., Morris, M., West, A., and Noden, P., (2005) Excellence in Cities: The National Evaluation of a Policy to Raise Standards in Urban Schools 2000-2003. Department for Education and Skills. Research Report RR675A.

Klungel, O.H., Martens, E.P., Psaty, B.M., Grobbee, D.E., Sullivan, S.D., Stricker, B.H.Ch., Leufkens, H.G.M. and de Boer, A. (2004) Methods to assess intended effects of drug treatment in observational studies are reviewed. *Journal of Clinical Epidemiology*, 57, 1223-1231.

Leow, C, Marcus, S, Zanutto, E, and Boruch, R (2004) Effects of advanced course-taking on math and science achievement: Addressing selection bias using propensity scores. *American Journal of Evaluation*, 25, 4, 461-478.

MacBeath, J., Kirwan, T., Myers, K., McCall, J., Smith, I., McKay, E., Sharp, C., Bhabra, S., Weindling, D., and Pocklington, K. (2001) The Impact of Study Support: A report of a longitudinal study into the impact of participation in out-of-school-hours learning on the academic attainment, attitudes and school attendance of secondary school students. Department for Education and Skills Research Report RR273. ISBN 1 84185 521 9

Morris, M. and Rutt, S. (2005) Excellence in Cities: Pupil outcomes two years on. Excellence in Cities Evaluation Consortium (NFER, LSE, IFS). Available at http://www.nfer.ac.uk/publications/pdfs/downloadable/MLM2005.pdf [accessed 24.7.09]

Morrison, K. (2009) *Causation in Educational Research*. Abingdon: Routledge.

Pan, W. and Frank, K.A. (2003) A Probability Index of the Robustness of a Causal Inference. Journal of Educational and Behavioral Statistics, Winter 2003, Vol. 28, No. 4, pp. 315-337.

Power, S., Whitty, G., and Wisby, E. (2006) The Educational and Career Trajectories of Assisted Place Holders. A report for the Sutton Trust, July 2006

Pritchett, L, (2002) It pays to be ignorant: A simple political economy of rigorous program evaluation. *Journal of Policy Reform*, 5, 4, 251-269.

Rosenbaum, P. R. (1986) Dropping out of high school in the United States: An observational study. Journal of Educational Statistics, 11, 3, 207-224.

Rosenbaum P R (1991) Discussing hidden bias in observational studies. *Annals of Internal Medicine* 115: 901–5

Rosenbaum, P. R. (2004) Observational Studies: Overview. *International Encyclopedia of the Social & Behavioral Sciences*, 2004, Pages 10808-10815.

Rosenthal, R, and Rubin, D.B. (1982) 'A simple, general purpose display of magnitude of experimental effect.' *Journal of Educational Psychology*, 74, 166-169.

Rubin, D. B. (2008) Comment: The Design and Analysis of Gold Standard Randomized Experiments. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 103, 484, 1350-1353.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., and Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Sewell, W.H. and Shah, V.P. (1968) Parents' education and children's educational aspiration and achievements. *American Sociological Review*, 33, 2.

Shadish, W., Cook, T. D., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Miflin.

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

Steele, FA, Vignoles, A & Jenkins, A. (2007) 'The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (3), pp. 801 - 824.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1998). Educational resilience (Laboratory for Student Success Publication Series No. 11). Philadelphia: Temple University Center for Research in Human Development and Education.

Winship, C. and Morgan, S.L. (1999) The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659-706.

Appendix

Equations for the simulation

We assume, without loss of generality, that all continuous variables Y, U, M , as well as RV1 to 5, are N(0,1).

Let RV1 be the 'common part' of M and U, and the 'distinct parts' be RV2 and RV3 respectively. In other words, for some constants, c and d,

$$M = c\ \mathrm{RV1} + d\ \mathrm{RV2}$$
$$U = c\ \mathrm{RV1} + d\ \mathrm{RV3}$$

Then

$$c^2 + d^2 = 1$$

and

$$c^2 = q_{um}$$

Then if

$$e_2 = f\,\mathrm{RV4}$$

Equation 2 becomes

$$Y = (\beta_0 + \beta_1 \mathrm{RV1} + \beta_2 \mathrm{RV2} + \beta_3 \mathrm{RV3} + \mathrm{RV4}$$

Calculating the correlations between this and M and U respectively, using the fact that all variables have mean 0 and variance 1, together with the knowledge that the RVs are mutually independent, gives a set of equations that can be rearranged to give

$$\beta_1 = \frac{rq - p}{q^2 - 1}$$

And

$$\beta_2 = \frac{pq - r}{q^2 - 1}$$

SPSS syntax for the simulation

This syntax runs in SPSS V15.0. The values of *r*, *p* and *q* in section 1. a) ii) should be set to the desired parameters before each run. For convenience, the same values can be entered into the final line so that a correctly labelled file is saved in the directory 'D:\TEMP\simresults' (the pathname can be changed as desired).

```
* simulation to show how a missing variable can make a lot of difference .

* Contains 2 macros:
1 to create a file of simulated data, save the required parameters in a dataset called 'combined'
2 to run the simulation repeatedly (50 times) and save the output .

set format=f5.3 .

*================================================
1. MACRO to run simulation once and output coeffs to SPSS dataset
using random values of correlation parameters, p, q, r, s
================================================ .

DEFINE !simulate_rand_once ()

  ********************************************************************************
  1.a) Set up the file
  ******************************************************************************** .

  * 1.a) i) create random variables ++++++++++++++++++++++  .
            new file.
            input program .
            loop n=1 to 100000 .
            compute rv1 = rv.norm(0,1) .
            compute rv2 = rv.norm(0,1) .
            compute rv3 = rv.norm(0,1) .
            compute rv4 = rv.norm(0,1) .
            compute rv5 = rv.norm(0,1) .
            end case .
            end loop .
            end file .
            end input program .
            execute .
            dataset name simulation .

  *1.a) ii)  compute random correlation parameters ++++++++++ .
            compute r = 0.3 .
            compute p = 0.7 .
            compute q = 0.3 .
            If ($casenum=1) #t = rv.uniform(0,1) .
            If ($casenum>1) #t = #t .
            compute t = #t .
            *this just sets all cases to the same random number .
            execute .

  *1.a) iii)  calculate coefficients ++++++++++++++++++++++++ .
            compute b0 = ( p*q - r)  /  (q**2 -1)  .
            compute b1 = ( q*r - p)  /  (q**2 -1)  .
            compute c  = sqrt(q) .
            compute d  = sqrt(1-q) .
            compute k1 = c*(b0+b1) .
            compute k2 = d*b0 .
            compute k3 = d*b1 .
            compute k4 = sqrt(1-k1**2-k2**2-k3**2) .

  *1.a) iv)  compute simulated variables++++++++++++++++++++ .
            compute M = c*rv1 + d*rv2 .
            compute U  = c*rv1 + d*rv3 .
            compute Y  = k1*rv1 + k2*rv2 + k3*rv3 + k4*rv4 .
            compute X = (t*U + (1-t)*rv5 > 0) .
            execute .
```

```
*******************************************************************************
1. b) Save outputs as 4 SPSS datasets:
RSQD, COEFFS, CORREL & RESULTS
******************************************************************************* .

*1.b) i)  Save assigned correlation parameters to 'combined' ++++++++++ .
            DATASET DECLARE combined.
            AGGREGATE
                        /OUTFILE='combined'
                        /BREAK=p
                        /q = MEAN(q)
                        /r = MEAN(r)
                        /t = MEAN(t) .

*1.b) ii)  Send regression output to 'rsqd' & 'coeffs' ++++++++++ .
            DATASET DECLARE rsqd.
            OMS
              /SELECT TABLES
              /IF COMMANDS = ['Regression']
                   SUBTYPES = ['Model Summary' ]
              /DESTINATION FORMAT = SAV
                          OUTFILE = 'rsqd'  .

            DATASET DECLARE coeffs.
            REGRESSION
              /MISSING LISTWISE
              /STATISTICS COEFF OUTS R ANOVA
              /CRITERIA=PIN(.05) POUT(.10)
              /NOORIGIN
              /DEPENDENT Y
              /METHOD=ENTER M X
              /OUTFILE=COVB(coeffs) .

*1.b) iii)  Calculate actual correlations and output to 'correl' ++++++++++ .
            DATASET DECLARE correl.
            OMS
              /SELECT TABLES
              /IF COMMANDS = ['Correlations']
              /DESTINATION FORMAT = SAV
                          OUTFILE = 'correl'  .
            CORRELATIONS
             /VARIABLES= Y M U X
              /PRINT=TWOTAIL NOSIG
             /MISSING=PAIRWISE .

            OMSEND .

*1.b) iv) Tidy up RSQD, COEFFS, CORREL  ++++++++++ .

            dataset activate RSQD .
            delete variables command_ to var1 .
            execute.
            rename variables (R=multR) .

            dataset activate COEFFS.
            select if (ROWTYPE_='EST') .
            execute.
            delete variables depvar_ to varname_ .

            dataset activate CORREL.
            select if (var2='Pearson Correlation') .
            execute.
            delete variables command_ to label_ var2 var3 .
            compute s_act=U .
            compute q_act=lag(M,1) .
            compute r_act=lag(Y,2) .
            compute p_act=lag(Y,1) .
            execute .
            select if (var1='X') .
            execute.
            delete variables var1 to X .
```

```
   ********************************************************************************
   1. c) Combine results together and format
   ********************************************************************************.
               dataset activate combined .

               MATCH FILES /FILE=*
                /FILE='rsqd'.
               EXECUTE.
               MATCH FILES /FILE=*
                /FILE='correl'.
               EXECUTE.
               MATCH FILES /FILE=*
                /FILE='coeffs' .
               EXECUTE.

               dataset close RSQD .
               dataset close COEFFS.
               dataset close correl.
               dataset close simulation .

               variable labels
                        p 'assigned correlation UNMEASURED with OUTCOME'
                        q 'assigned correlation UNMEASURED with MEASURED'
                        r 'assigned correlation MEASURED with OUTCOME'
                        t 'assigned strength of relationship of GROUP with UNMEASURED'
                        p_act 'actual correlation UNMEASURED with OUTCOME'
                        q_act 'actual correlation  UNMEASURED with MEASURED'
                        r_act 'actual correlation MEASURED with OUTCOME'
                        s_act  'actual correlation  UNMEASURED with GROUP'
                        multR 'multiple correlation MEASURED and GROUP with OUTCOME'
                        CONST_ 'Intercept'
                        M 'coefficient of MEASURED'
                        X 'coefficient of TREATMENT GROUP'  .

               variable width p to X (6) .


!ENDDEFINE .




*================================================
2. MACRO  to repeat simulation, paste coeffs into RESULTS and save
================================================ .

DEFINE !repeat_sim (reps=!TOKENS(1)
   /session=!TOKENS(1) )

   !simulate_rand_once .
               dataset name results .

   !DO !I = 2 !TO !reps .
               !simulate_rand_once .
               dataset activate results .
               ADD FILES /FILE=*
                        /FILE='combined'.
               EXECUTE.
               dataset close combined .
               output close all .
   !DOEND .

   SAVE OUTFILE=!CONCAT("'D:\TEMP\simresults\random_session",!session,".sav'")
                /COMPRESSED.

!ENDDEFINE .


 !repeat_sim reps=50 session=373 .
```