*From cross-sectional multi-level modelling to longitudinal*

*analysis of country-level variables*

A review of existing research

**David Bartram**

**Patrick White**

University of Leicester

December 2022

Project website: https://le.ac.uk/mcs/about/research/health-communication-and-wellbeing/from-cross-sectional-multi-level-modelling-to-longitudinal-analysis

# Executive summary:

For social-science research asking about the impact of a country-level independent variable (e.g. inequality) on some aspect of individual-level experience, most researchers use cross-sectional multilevel models (MLM). A cross-sectional analysis in general is understood to carry risks of bias, especially for conclusions framed in terms of causal relationships. Most researchers are aware of these risks. But the prospects for constructing a longitudinal analysis are commonly held to be very low.

We report results of our own investigation into the persistence of cross-sectional approaches, exploring the reasons researchers give for their choices. Most researchers believe that the data required for a longitudinal analysis are lacking. In our view, a longitudinal analysis is often possible, via use of repeated cross-sectional survey data (e.g. the European Social Survey). An analysis using the approach we suggest does not appear as an available option to many researchers – and our research indicates that 'lack of data' is the main reason. It appears that this view comes from the belief that a longitudinal analysis would require cross-national *individual-level* panel data – a form of data that is indeed lacking. We explore the possibility that implicit assumptions are feeding this view – including assumptions about the need for particular kind of control variables. Our core conclusion is that a more effective analytical approach for questions of this sort is likely available in many instances. The path to improvement along these lines consists of paying explicit attention to core components of the analysis where at present certain features are typically the outcome of implicit decisions.

# Table of Contents

# Introduction

This report evaluates the state of current research for research questions that have a specific structure. Researchers in the social sciences are commonly interested in evaluating whether there is a causal relationship between a country-level independent variable and an individual-level dependent variable. (An example is: the impact of inequality on life satisfaction.) Currently, the overwhelming majority of research exploring questions with that structure is conducted in a cross-sectional mode. We articulate a framework that enables application of a longitudinal analysis – an approach that has clear advantages, especially if we want to interpret results in causal terms.

The purpose of this report is to assess existing research in ways intended to reveal reasons for the persistence of prevailing approaches. Quantitative researchers generally know about the limitations of cross-sectional analysis – in particular, the risks of drawing causal conclusions from analysis of that sort. The question then becomes: why does a cross-sectional approach persist? How can we understand why researchers do not adopt a longitudinal framework? Answers to those questions will help us understand what we might do to encourage and facilitate a more effective engagement with questions having the indicated structure.

The core idea motivating our project is that, while many researchers appear to perceive a 'data problem', the real obstacle consists of unexamined assumptions about the control variables needed for the statistical models we would construct in this context. In many instances researchers apparently believe that an effective model requires a large number of control variables, including many variables pertaining to individual-level characteristics ('level-1' controls). We are confident that these variables are not in fact necessary as controls. Once that point is clear, the 'data problem' ostensibly inhibiting a longitudinal analysis is mitigated to a great extent. Our core claim, then, is that there is inertia in this sense in part because of taken-for-granted ideas about what sort of control variables are needed for research questions of this type.

This report presents some findings of an investigation into the research practices evident in published articles, an investigation where we consider existing research in these terms. We also carried out a brief on-line survey inviting authors to articulate their views on core topics pertaining to analytical choices. We describe our methods for carrying out this research in the relevant sections below.

To proceed, we first describe our own longitudinal approach in more detail. We then consider potential reasons why an approach along these lines has not previously been adopted. (We do not confidently assert that we are the very first researchers to develop the core idea – but the question would then become: why did it not take root?) Subsequent sections report the findings from the research carried out specifically for this report, where we seek to gather data that might substantiate (or indeed dispel) our hunches.

# A longitudinal framework for investigating country-level questions

Suppose we want to know how economic inequality might affect people's life satisfaction. A common sense hypothesis might suggest that inequality generally makes people less satisfied with their lives; a small number of wealthy people might like it, but the larger number of poorer people would dislike it.

How could we know? The approach prevailing in existing research is to construct international comparisons, using multi-level models (MLM). The typical finding (e.g. Kelley & Evans 2017) goes against common sense: especially in poorer countries, average life satisfaction is *higher* in more unequal countries (while in wealthier countries there is no particular relationship). These results are best described as associations, rather than causal effects. Perhaps life satisfaction is higher in more unequal countries, but this does not mean that life satisfaction is higher there *because* those countries are more unequal.

To know whether higher inequality *leads to* higher life satisfaction, it is much better to investigate how life satisfaction changes over time, in line with changes in inequality. A cross-sectional analysis gives results that are vulnerable to 'omitted variable bias'. The cross-sectional results could lead us astray, exactly in the sense articulated just above: life satisfaction might higher in more unequal countries, but this does not mean that life satisfaction is higher there *because* those countries are more unequal. The cross-sectional association might reflect the influence of other important differences between countries. A longitudinal analysis, investigating change over time, is more effective in isolating the impact of those differences. In statistical terms, a longitudinal analysis (constructed in a 'within' mode, a.k.a. 'fixed effects') nullifies the influence of country-level differences that don't change over time.

The focal independent variable in this context is inequality – a country-level variable. The dependent variable is life satisfaction – an individual-level variable. To give a foundation for answering the research question, we need to include multiple countries in the analysis. What we then see is that there are no cross-national panel data at the level of individuals. What we propose, then, is to use data from repeated cross-sectional surveys (e.g. the European Social Survey). We use the individual-level data on life satisfaction from those datasets to construct country-level averages of life satisfaction for each survey wave. These country-wave averages thus take the form of panel data at country level. These data can then be matched to time-corresponding values of inequality (e.g. Gini coefficients), perhaps with a lag that reflects our sense of causal direction. We now have the essential data required for a 'within' analysis. We can also include time-varying country-level variables that might be relevant as controls.

We can then ask: does adoption of a longitudinal framework make a difference to the results and the conclusions we draw? In this instance it does. Cross-sectional research (as described above) gives a 'positive' association: life satisfaction is higher in more unequal countries (especially when considering 'developing' countries). Our own findings (in a manuscript currently under review) go in the opposite direction: as inequality increases over time, life satisfaction generally *decreases*. The decrease is larger when we investigate wealthy/developed countries – but even in 'developing' countries the result is 'negative'. The risks of omitted variable bias in cross-sectional research are

quite real in this example. Adopting a longitudinal framework is highly consequential – and we have good reasons to prefer the longitudinal results.


## Why would a longitudinal analysis seem impossible/infeasible?

We suggested above that researchers might not discern feasibility for the approach we describe because of a 'data problem'. As noted, to evaluate the causal proposition in general terms, we need to investigate more than one country (ideally, using data from a large number of countries). To construct a 'within' analysis, we need panel data. What is lacking, then, is panel datasets that cover more than one country. More precisely, there are no *individual-level* cross-national panel data. Panel datasets with individual-level data are invariably *national* entities (e.g., the *British* Household Panel Survey, along with its successor 'Understanding Society'). Insofar as researchers articulate a reason for conducting cross-sectional analyses, we would expect to see that this 'data problem' constitutes the core reason: we simply lack the data required to adopt a longitudinal approach.

In the previous section we identified a way around the data problem for the dependent variable (life satisfaction): use individual-level data from repeated cross-sectional surveys to construct country-level panel data, via aggregation. But a credible analysis likely needs more information, beyond the dependent and focal independent variables. To address the possibility of bias, even a longitudinal model is likely to need control variables. Many researchers are likely to believe that an analysis where life satisfaction is the dependent variable will require individual-level variables as controls. The 'data problem' exists beyond the dependent variable, it seems.

The view that individual-level variables are needed as controls comes from a misconception, in our view. Many researchers pay little attention to the question of how to select control variables effectively. In some instances, a methods section addresses this angle by appealing to precedent: in a previous investigation, someone else included the following controls, so that's what I am going to do as well.

When we see articulation of a criterion for selecting control variables, the criterion typically takes the following form: the needed controls are 'other determinants' of the dependent variable. So, with life satisfaction as the dependent variable, we need to ask: what else influence people's life satisfaction? That question leads us to a list that can become quite long: age, sex, education, income, employment situation, partnership situation, religiosity, friendship networks… In more general terms: our target quantity is X→Y. We use 'W' to designate control variables. So, the 'other determinants' criterion is summarised as: control for W where W→Y.

The purpose of control variables is to address the possibility of bias in our estimate of the target quantity (here, the influence of inequality on life satisfaction). With that purpose in mind, the 'other determinants' criterion is incomplete, and in some instances it might lead to use of controls that exacerbate bias rather than remedying it. Lack of clarity on this point amounts to the 'unexamined assumptions' explanation for why some researchers persist in using a cross-sectional analysis.

To mitigate bias, we need (as controls) variables that are *antecedents* of the focal independent variable, in addition to being 'other determinants' of the dependent variable. In other words, to get an unbiased estimate of X→Y, we need controls where W→Y *and* W→X. This criterion is well established, e.g. in Morgan & Winship (2007) and Gangl (2010). We need to ensure that we do not include controls that are *influenced by* the focal independent variable (X→W). If we include controls where that pattern prevails we will induce bias in our estimate of X→Y (see Angrist & Pischke (2009) on 'bad controls' and Rohrer (2018) on 'overcontrol bias').

In connection with W→X, we now need to ask: starting with the 'other determinants' of life satisfaction, are there any variables that are reasonably considered antecedents of inequality? For individual-level variables that influence life satisfaction, the answer is very likely: no. The level and/or trend of inequality in any particular country is unlikely to be influenced by individuals' education, their religiosity, their friendship networks, etc. These variables might seem relevant as controls insofar as they amount to W→Y. But they are, in the end, not relevant, because they do not *also* amount to W→X. Including them in our model is unlikely to have any substantial impact on our estimate of X→Y.

Where we see researchers including individual-level controls in their models for research 'in scope' here, then, we will conclude that they are doing this for reasons that amount to 'unexamined assumptions' about control variables. This practice might come from following precedent, without articulating a selection criterion for controls – or it might come from articulating an *incomplete* (i.e., incorrect) criterion. These reasons constitute unnecessary obstacles to adoption of a longitudinal analysis, because they exacerbate the appearance of a 'data problem'. Even if the possibility of using aggregation for the dependent variable was apparent, the 'need' to include individual-level controls would likely function as an impediment to the prospects of constructing a longitudinal analysis.


## Assessment of existing research

The ideas described in the previous section were formed from our previous experience of reading published research addressing questions 'in scope' (i.e., for having the indicated structure). This report evaluates whether these ideas are borne out in a more extensive consideration of studies, extending beyond our own substantive research interests. A multi-disciplinary engagement (going beyond sociology) is especially necessary. The core questions are:

- How common is it for researchers to address such questions in a cross-sectional mode?
- What reasons (if any) are given for the adoption of a cross-sectional analysis? Is a longitudinal approach contemplated? If yes but that alternative is then rejected, why is it rejected?
- How do researchers motivate their decisions about selection of control variables? Do these decisions play a direct role in the adoption of a cross-sectional analysis?

These questions were addressed via research in two modes: reading published articles, and gaining responses to a brief on-line questionnaire asking researchers to disclose their general ideas connected to these issues. The answers from this research can be summarised succinctly:

- Longitudinal analyses for questions of this sort are rare. Cross-sectional analyses are pervasive, especially for questions requiring international comparisons.
- In most instances, no particular reason is given to justify the adoption of a cross-sectional approach. Occasionally the question is raised – and in those instances there is commonly a brief statement indicating that a longitudinal analysis is not possible because the required data are lacking.
- It is very rare for researchers to give justifications for the selection of control variables. In many instances this topic is addressed in a single paragraph that only gives the list of selected controls, without providing reasons. Where a justification is given, 'precedent' is the typical answer.
- We did not identify a single instance where the research connected their cross-sectional approach to the 'need' for individual-level control variables. (In general, the topic of control variables receives *very* little direct attention in most quantitative empirical work.) In this sense many researchers appear to be taking a great deal for granted about how quantitative research for this sort of question should be conducted.

For the most part, then, our expectations (based again on reading articles connected to our own narrower substantive research interests) were borne out by the research conducted specifically for this project.

## Supporting evidence – on-line survey

We wrote to 52 researchers inviting them to respond to an on-line survey. The main mechanism for inclusion was: they were authors of the published articles identified by a research assistant (for the second component of our research, described below). We received usable responses from 15 people. Of these, ten identified themselves as sociologists; there were single responses from scholars in health, public health, epidemiology, and economics, as well as combinations of these. Length of post-PhD career ranged from two years to 40 years, with no particular concentration within that range. Eleven respondents gained their PhDs from a university in a European country (including the UK); four identified the USA or Canada for this question. Our identification of potential respondents did not use any sort of mechanism that would enable us to assert that our 'sample' is representative of some broader population.

We asked questions on four topics pertaining to their use of multi-level models: 1) Are you mainly interested in evaluating causal effects, or 'associations'? 2) What is the 'level' of your focal independent variable? 3) Did you explore the possibility of conducting a longitudinal analysis – and if not, why not? 4) How did you decide which control variables to include (and exclude)? The questionnaire for that survey is provided in an appendix below.

Three of the 15 respondents said that their intention was to investigate a causal relationship. Several others said that they were interested in the possibility of a causal relationship but understood that their data would not enable them to interpret their results in those terms. Two suggested that a causal relationship was indicated by their theoretical framework but again the data did not enable them to draw causal conclusions. Several respondents said simply that their intentions were limited to identifying 'associations' or 'relationships'.

We approached potential respondents on the basis of having identified their use of multi-level models in a cross-sectional mode. We then asked: did you consider a longitudinal analysis – and if this did not seem possible, why not? Several respondents did not answer in line with our premise: they wrote about use of a longitudinal approach, likely with a project in mind that was different from the published article we had initially consulted.

For the respondents who *did* discuss use of a cross-sectional approach, the answer to our question was exactly in line with our expectations: a longitudinal analysis did not seem possible because of lack of data. (For example: 'the data set was cross-sectional'; 'the dataset was not a panel dataset'; 'no data available'.) The most interesting response was: 'This was a repeat cross-sectional dataset. Of course longitudinal would be preferred but data unavailable.' This answer appears to suggest that a longitudinal approach *would* have been possible, via the approach we described above. The assumption here appears to be that the research would have required *individual-level* cross-national panel data.

Responding to our questions about selection of control variables, most respondents referred to 'precedent' or 'previous research'. Several respondents referred to 'theory' – an answer we think equates to 'previous research'. Other answers invoked 'availability of data'. One respondent indicated use of a 'casual DAG' (i.e., directed acyclic graph) – the only response that suggests (quite strongly) use of the right criterion, in line with the perspective we describe above. Another response that seems in line with that perspective used the words 'control, confounders, exposure, and outcome' – words that are commonly used in conjunction with DAGs. A third response that might be read as in line with that perspective used the phrase 'variables that confound the relationship of investigation'.

We asked specifically about decisions to *exclude* potential controls. Many respondents had little to say in response to that question. Several responses referred to 'statistical' angles, e.g. multi-collinearity, (lack of) significance, and lack of association with the main variables of the study (with the control then dropped for the sake of 'parsimony'). One referred to 'demands of reviewers'. There were no responses to this question that made connections to the 'levels' of the study – i.e., no exploration of the possibility that individual-level controls might not be required for questions asking about the impact of country-level variables. In general, it seemed that most respondents thought more about the basis for including controls, as against any basis for *excluding* controls.

We recognise that some of our respondents might have ideas about these topics that are more developed than their responses to our questionnaire suggest. We promised potential respondents that our survey would be 'brief' and would not require much of their time. It would be unsurprising if some respondents provided answers that did not elaborate on a topic that they fully understand is more complex than what is apparent from their questionnaire responses.

Even so, the data we have reinforce our sense that there are some missed opportunities in some of the existing research we have for questions that are 'in scope' for our project. Many researchers are clearly interested in the possibility of identifying causal relationships. Even the ones who say they are interested in 'associations' might want to explore causal relationships; it seems likely that they say they are interested in associations in part because they perceive (even if they do not say) that the data they believe are required are simply not available.

## Supporting evidence – review of published studies

For this component of the project we engaged a research assistant. The RA was asked to conduct a search (using Google Scholar) to identify published research that falls 'in scope' (in the sense that the studies investigate a research question asking about the impact of a 'Level 2' variable on a 'Level 1' outcome, thus using some form of multi-level modelling). The RA was then asked to describe each study with reference to the following questions/fields:

- Data (source)
- Analytical approach, i.e., cross-sectional vs. longitudinal
- Identification of key variables, at different levels
- Control variables
- Justification for selection controls (if any)

The RA compiled a spreadsheet summarising 68 publications in these terms, drawing on a range of social science disciplines (including sociology, geography, criminology, health, politics, finance, education, and transport). This 'sample' of publications was not selected by any sort of randomising mechanism; it can be considered indicative rather than representative.

Our overall summary of findings on p. 6 above comes in part from consideration of the RA's data, derived from her evaluation/description of the publications she identified. These data reinforce the following points: 1) Most studies using multi-level modelling were cross-sectional, especially in instances where the focal independent variable was a 'national' entity of some sort and the research therefore required international comparison. 2) Most studies included a wide range of individual-level control variables. 3) In many instances there was no justification offered for the inclusion of these controls; if a justification was offered, it usually took a form suggesting that we need to control for 'other determinants' of the dependent variable.

These descriptions apply across the range of disciplines covered by the studies identified by the RA. In other words, we do not see reasons to believe that researchers in some disciplines are more likely to devote more attention to the topic of control variables. Judging from what appears in published research, quantitative social scientists in general (though by no means universally) seem to take for granted a great deal about the function and selection of control variables. In our view, the implicit ideas many people seem to have operate as an obstacle to adoption of a longitudinal analysis: researchers appear to believe that they need individual-level control variables in their models, a belief that runs up against the absence of cross-national individual-level panel data.

## Summary

The perspective we describe in this report, together with the evaluation of existing research, supports the contention that there is an opportunity for researchers to move beyond cross-sectional multi-level modelling when addressing research questions that have the indicated structure. When investigating the impact of a country-level variable on an individual-level outcome, a longitudinal analysis merits exploration: if there is repeat cross-sectional data for the dependent variable, then aggregation to country-level means can be the foundation for a country-level panel analysis. The use of a 'within' specification then offers a basis for overcoming the limitations of a cross-sectional comparison. Any bias from omission of individual-level control variables is likely to be small in comparison to the bias avoided by the 'within' model's effectiveness in controlling for time-constant country-level confounders. The results from an analysis along these lines likely offer a stronger foundation for conclusions expressed in causal terms.

## Cited works

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology*, *36*(1), 21–47.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.

## Appendix: on-line survey questionnaire

1.  Thinking of your recent work using MLM, was the research question focused on investigation of a (potential) causal relationship? If not, how was the research question framed? (e.g. an 'association'? a 'relationship'?) [For this and subsequent questions, it is likely best to focus on one particular study (in case you have used MLM in multiple publications).]

2.  If a causal relationship was being investigated, did the work involve a research question asking about the effect of a 'level 1' variable? Or was the focus on the effect of a 'level 2' variable? In other words: could you please tell us whether the analysis was focused on the effects of a *specific* independent variable and at what level that variable was located?

3.  MLM investigations are typically cross-sectional. Did you explore the possibility of constructing a longitudinal analysis? Did it work? If a longitudinal analysis did not seem possible, what was the obstacle?

4.  How did you decide which control variables were needed for the model? Was there a specific/explicit criterion for inclusion of controls? Or, was it perhaps a matter of following precedent in previous research?

5.  Did you make explicit decisions on exclusion of any potential control variables? In other words, were there variables you considered as potential controls but then you decided not to include them? What was the basis for decisions of that sort?

6.  What is your main academic discipline?

7.  How many years have you been working as an academic researcher (post-PhD)? (An approximate answer is acceptable.)

8.  In which country did you gain your PhD?