‘Young Lives’ Peru modeling Project, 2016 – 2017

Preliminary report : structural equation modeling of Younger Cohort educational outcomes

Colin Tredoux and Andy Dawes, December 2016

==================================================================================

# Overview

One of the questions that Young Lives seeks to answer is: *how, and at what points in the early life course does poverty and associated risks shape adolescents’ and young people’s well-being and developing capabilities?*

Young Lives longitudinal data and large sample size permit exploration of this question using Structural Equation Modelling (SEM) - one possible multivariate analysis technique that enables us to test a model of the manner in which selected variables predict an outcome of interest. In this output we report on SEM using data from the Peru Younger Cohort (YC) to predict three outcomes at 12 years of age (Round 4): expressive language, mathematics achievement, and grade for age. Education outcomes such as these are key indicators of children’s developing capabilities and are associated with later educational performance. Understanding of the factors that influence educational achievement in children affected by household poverty and associated risks has implications for policy.

We have chosen the YC as these children were enrolled between 6 and 18 months of age and followed up at five, eight and twelve years. It is well established that the early childhood period is important in shaping development in the long term. Availability of data for the early childhood period in the YC enables us to explore its influence together with those occurring later during middle childhood – a period during which family influences and support for learning in school becomes influential. OC children were enrolled at eight years so the early period cannot be examined in this cohort.

It must be stressed that our analysis is constrained by the availability of appropriate variables. We are aware that educational outcomes are a function of a complex interaction between factors endogenous to the child, and home and school influences. In this analysis we have been restricted to the use of the household surveys and child questionnaires and tests of language and numeracy. An obvious missing link in our analysis is the effect of the school on the outcomes of interest. A number of the cohort children participate in the school surveys, but limiting the SEM to children for whom cohort and school survey data was available would have significantly reduced the sample available for analysis and compromised the representativeness of the cohort. We also explored the potential of the 2009 Community Survey to provide indicators of school quality by linking YC children to the places and schools for which key informants provided relevant information. This was not successful, as one could not be sure of a valid link between the child and the educational institution, and further, there was a considerable number of missing records on indicators of interest. The contributions of schooling are therefore best examined separately through the Young Lives school effectiveness surveys. A further possibility worth exploring is a sub-study of those children who participated in both the cohort and school effectiveness studies to examine the relative contributions of school and home to educational outcomes of interest.

# Identification and selection of outcome variables

We accessed YL data sets for Rounds 1 to 4 through the portal at ukdataservice.ac.uk, including the ‘constructed’ data set for all four rounds (http://dx.doi.org/10.5255/UKDA-SN-7483-2).

Data dictionaries and questionnaires for the YL project, Rounds 1 to 4, were examined, and a preliminary set of outcome variables was identified for potential secondary analysis. Reconciliation of the data dictionaries with questionnaires and data sets reduced the outcome variables to five at Round 3, and three at Round 4. In the analysis we present here, we consider only the outcomes variables common to Rounds 3 and 4. The variables were as follows: **Expressive language** (Peabody Picture Vocabulary Test - raw score), **Mathematics achievement** (Mathematics Achievement test – percentage score), and ‘**Years over age for grade**’(computed by assuming that children turn 7 in the year that they start school in Peru, and computing the grade they would be in with successful promotion at each grade)[[1]](#footnote-1). The three variables were combined in the modelling exercise as a latent variable; that is, we do not model each of the variables separately, but rather as an aggregate. Two of the outcome variables (expressive language and mathematics achievement) were themselves the sum of a great many individual items (over 100 for the PPVT alone), but we decided to consider only the summed variables themselves, rather than introduce much greater complexity to the model.

The inclusion of outcome variables at Rounds 3 and 4 suggests the possibility of treating them as longitudinal elements in the design. However, they are constituted slightly differently at the two rounds, and are very strongly correlated, which would likely make for greater complexity in the analysis, and we opted instead to treat the outcome at each round as a separate test of the model. This is not a strong cross-validation test, but it is a useful check on what is quite a complex model.

# Identification and selection of predictor variables

In order to find and select appropriate predictor variables for the modelling exercise, we followed much the same process outlined above for outcome variables i.e. a thorough examination of the questionnaires, data dictionaries, and data files. We ended up with 38 predictors, which in our model reduced down to 9 latent variables. In the model testing we pruned these down further to 6 latent variables, constituted by 29 manifest variables (the variables as originally measured).

Table 1, below, summarises the variables we selected, with corresponding hypothetical latent variables.

**Table 1**. Predictor variables used to model educational outcomes in the younger cohort of the YL data set, Rounds 1 to 4

|  |  |
| --- | --- |
| **Latent variable** | **Item(s) – manifest variables, and round at which measured** |
| Caregiver status  Caregiver literacy  Economic well-being    Investment in education  Caregiver’s (mother’s) psychological well-being    Child’s health | Two items (partner status, and mother’s age at birth). Round 1  Two items (Mother’s years of education (re-scaled), and mother’s literacy (binary: literate or not)). Round 1  Three items (wealth index (see peru constructed files), food insecurity, and family expenditure). Averaged over Rounds 1 and 2.  Four items (hours index child spends doing chores, school type index child attends (private or not), hours child spends studying per day outside of school, parental encouragement of reading at home). Round 3  Fifteen items (selected from larger set of 20) – subset of the SRQ 20. Round 1  Three items (stunting over rounds 1 to 3 [recoded as 6 point scale], long term problems with vision and/or respiration. Averaged over rounds 1 to 3. |

# Preliminary model definition

It is conventional to represent SEM models diagrammatically; we do so here in a simplified manner. That is, we show the predicted connections between latent variables (the regression component of the structural model), but not the measurement component of the model. The measurement component should be sufficiently clear from Table 1, and the description of the outcome variables.

Figure 1 shows the predicted regression model. Note that we constructed this model on the basis of i) prior knowledge of the literature, and ii) availability of appropriate measurement variables in the YL datasets. Some predicted paths are on a much surer footing than others. There is no doubt that we have had to be somewhat eclectic and opportunistic in our approach.

1 yr 5yrs 9 yrs 12 yrs

Round 1

Round 2

Round 3

Round 4

**Figure 1**. Proposed regression component of SEM for predicting education outcomes

# Testing the measurement component of the model

As a precaution, we tested the measurement component of the model prior to inclusion of the regression paths: if the measurement component is not satisfactory, the full structural model is unlikely to fit the data well. We conducted a confirmatory factor analysis with the R Programming Language for Statistical Analysis, and the R package LAVAAN (for latent variable analysis). The analysis was conducted iteratively, adjusting the measurement model where problems of correlated error terms or analytic feasibility were discovered. This led to i) the exclusion of five of the 20 items from the SRQ (the measure of mother’s mental health), and ii) the exclusion of two of the school environment variables (perceived school quality, and reported prevalence of punishment at school). In the former case, keeping the five excluded items would have meant incorporating correlated error terms, which we preferred to avoid. We are satisfied, nonetheless, that the remaining 15 items capture the construct of interest very well. In the latter case, the combination of high levels of missing data and inter-correlatedness of variables made the analysis infeasible without additional, restrictive conditions being imposed on the model.

The overall degree of fit of the measurement model was good: the commonly computed measures of model fit RMSEA, sRMR, and CFI all showed an acceptable degree of fit, and for RMSEA and sRMR the fit was very good (RMSEA = 0.036 [95% CI = 0.034; 0.038], sRMR = 0.035, CFI = 0.91).

# Testing the full structural model (measurement + regression)

After adjusting the measurement component of the model, as described above, we tested the full structural model, again using R and LAVAAN, and using the Round 4 education outcome variable. We also did this iteratively, and had to remove a single indicator latent variable reflecting whether the language of instruction at school matched the home language of the index child (the error term correlated with the error terms of several other latent and manifest variables). We found that the model constituted a fairly good fit to the observed data, except that many of the paths were not statistically significant. This is a good result rather than a bad result, though, as it simplifies the model considerably, and also shows important mediating relationships in the model. The model we accepted is shown as Figure 2, with standardized path coefficients.

RMSEA = 0.028 [0.026; 0.31]

CFI = 0.95

sRMR = 0.031

Total N = 1643

R2 = 0.72

0.85

0.60

0.91

0.82

0.22

1 yr 5yrs 9 yrs 12 yrs

0.344

Round 1 12 months

Round 2

Round 3

Round 4

**Figure 2**. Accepted regression component of SEM for predicting education outcomes in Round 4.

Especially important in this diagram is the fact that many paths are not present. In places this indicates mediating relationships while in others it indicates the absence of a variable from the causal chain (e.g. caregiver’s psychological well-being). Of particular interest is that the effect of economic well-being on education outcomes is completely mediated by investment in education which is made up of three items: the hours the index child spends doing chores and in studies, whether or not the child attends a private school, and whether or not parents encourage reading. All coefficients are standardised, and are significant at p < 0.001. The model explains 72% of the variance in the outcome latent variable Once could reasonably speculate that a strong measure of school quality could improve the model further..

To some degree, our analysis is at the mercy of sampling variation – regression methods are known to capitalize on chance, and not to generalise particularly well. It is common to test this, albeit imperfectly, with some form of cross-validation. We decided to test the model, as explained earlier in this overview, by cross validating it on the Round 3 education outcome latent variable. This is not a strong form of validation, but it does constitute a test of the model. The model showed very similar results for the Round 3 outcome variable (RMSEA = 0.029, sRMR = 0.032, CFI = 0.95, overall R2 = 0.70). All the regression paths remained statistically significant.

# Comparing the model across poverty groups

The model we report on above was tested on the entire sample, but there is an argument for testing its robustness across groups within the sample that differ substantially in terms of their relative poverty. In the Peruvian context it seems appropriate to contrast groups that live in rural rather than urban areas, and that are indigenous (in the sense of speaking an indigenous language at home, or of having at least one parent whose home language is indigenous) rather than Spanish. We therefore created a variable that reflected this division within the sample, and tested the model referred to above on the groups split in this manner.

Whereas the model provided a good fit for the urban, non-indigenous language group, it was not a good fit for the rural, indigenous sample. We therefore revised the model using standard methods (inter alia modification indices, residual error estimates), and adjusted it accordingly. This post hoc adjustment resulted in a quite different model, which we show as Figure 3 below.

0.28#

0.29✝✝

RMSEA = 0.033 [0.022; 0.43]

CFI = 0.91

sRMR = 0.05

Total N = 344

R2 = 0.58

0.52

0.58

0.51

0.17

0.22

Round 1

Round 2

Round 3

Round 4

**Figure 3**. Regression component of SEM for predicting education outcomes in Round 4 for *indigenous, rural children*. All coefficients are standardised, and are significant at p ≤0.05, except ✝, where p ≤ 0.11 and #, where p < 0.06

For indigenous rural children. Family economic well-being from Round 1 through Round 2 (the early childhood period) directly influences education outcomes at age 12. The influence is also mediated by investment in education when the child is 8 years old. This indicates that the effects of the family economic situation on education outcomes is explained by the time the index child spends doing chores and in studies, whether or not the child attends a public or a private school, and whether or not parents encourage reading. The effects of maternal psychological well-being are evident but again, this is mediated through investment in education.

The model for indigenous rural children explains less of the variance in education outcome, in comparison to that for the entire sample. It is not clear why this is the case. Nevertheless, it is a reasonably successful model, and the amount of variance in outcome scores that it resolves (58%) is fairly high.

We also ran the model for boys and girls separately, and found very similar results : all the regression coefficients were significant for both groups, and of similar size. We did not test for invariance of the model across the grouping factor, since this was not an a priori prediction. Although we wanted to examine the model at additional sub-grouping levels, especially across gender within the rural, indigenous sample, this would have meant testing quite a complex model with sample sizes that are likely too small for acceptable levels of statistical power.

# Summary, and limitations

We developed and tested a structural equation model of education outcomes in Peru, using data from the Younger Cohort in the longitudinal Young Lives data set. This involved identification of a suitable set of outcome and predictor variables from questionnaires and data dictionaries, and a model combining the predictors both cross-sectionally and over time. The model needed minor refinements once tested, and suggested a pattern of causal relations between caregiver literacy (at Round 1), economic well-being (at Rounds 1 and 2), parental and child investment of time, and resources in education (at Round 3), and education outcomes (language, mathematics, and successful grade progression; at Round 4). The degree of model fit was good, and a cross-validation of the model with educational outcomes at Round 3 showed a very similar degree of fit. However, whereas the model was a good fit for Spanish speaking children in urban areas, it was not a good fit for indigenous children in rural areas. A separate, post-hoc model for the latter group suggested a more complex model, in which maternal psychological well-being and economic well-being played a greater role.

It is important to point to some limitations of the models discussed here. Firstly, all structural equation models make assumptions about the direction of causation, and if the assumptions are wrong, the modeling results may also be wrong. In the present case we attempted to use the longitudinal nature of the data to guide the model logic, but this is of course open to contestation. In further work, we propose to compare models that make contrasting assumptions about the direction of causation for the predictors in the extant model. But more generally, the model we proposed for testing is open to contention – we constructed a model that made sense to us, given our knowledge of the literature, but otherw may have different views.

Secondly, structural equation models can only be as good as the variables measured and made available for analysis. If important variables have been omitted, the models could be wrong, or weaker than would have been the case. This is not something we have much control over, given that we are conducting a secondary data analysis.

# Next steps

The exercise reported here has been in the nature of an experiment to test whether or not SEM is feasible and fruitful for analysing YL data. It does appear to be viable, and may offer some insights into the complex determinants of education outcomes.

A key question is whether we repeat this analysis on YC for Round 4 outcomes in other countries, or wait until we have Round 5 data, which will permit us to examine the effects of changes during adolescence on educational outcomes at 15-16 years of age. This would seem to be the wiser route.

Regarding the older cohort, as already indicated, we do not have data from the early childhood period but we would be able to examine outcomes at 19 using Round 4 data. We recommend that we proceed with the development of a model to test influences on outcomes at 19 for the Peru Older Cohort. This will enable an examination of their transitions to work or further education and training. If this is fruitful, we can proceed to examine predictors of these transitions in the OC in the other three countries.

1. Please note that the constructed data files made available on the Young Lives data site (e.g. peru constructed.sav) appear to contain an error in the variable reflecting child’s grade in Round 3 – the grades appear to be one grade lower than they should be. We recovered the information from one of the original survey data files instead. [↑](#footnote-ref-1)