A Model for Quality of Schooling

Massoud Moussavi and Noel McGinn
Causal Links, LLC
Chevy Chase, Maryland
moussavi@causallinks.com, mcginn@causallinks.com

Abstract

A key challenge for policymakers in many developing countries is to decide which intervention or collection of interventions works best to improve learning outcomes in their schools. Our aim is to develop a causal model that explains student learning outcomes in terms of observable characteristics as well as conditions and processes difficult to observe directly. We start with a theoretical model based on the results of previous research, direct experience and experts' knowledge in the field. This model is then refined through application of supervised learning methods to available data sets. Once calibrated with local data in a country, the model estimates the probability that a given intervention would affect learning outcomes.

Introduction

There is a large research literature in education that describes "what matters" for learning outcomes. For example, research shows that learning is affected by teacher quality, time spent outside classrooms on learning, and textbooks. The findings typically are based on regression analyses and experimental studies.

The research results are the basis of questionnaires and protocols designed to collect key information about how schools are doing. Most surveys collect information on a number of "input" variables (school facilities, organization), teachers (their academic education, training, method of teaching), students (attendance rate, time spent on homework, health, nutrition), and students' family (social status, where they live).

Policy analysts typically focus on input variables. Much of the policy research is based on the "production function" approach, in which inputs such as school physical facilities, family attributes, teacher attributes are linked to student achievement. While these efforts have contributed to the understanding of the factors associated with student learning, some argue that research on education production functions simply has not shown a clear, systemic relationship between resource inputs and student learning outcomes. [Hanushek, 2008]

This is because the effect of input variables on learning outcomes often is mediated by contextual variables such as time spent learning outside classrooms, curriculum coverage, teacher skill, teacher motivation, student motivation and student engagement attention. Our aim is to develop a causal structure that includes such variables and their inter-relationships and explores their effects on the learning outcomes. Over the last two decades a significant body of research has focused on developing mechanisms for learning Bayes net and causal structures from data. [Heckerman 1998, Pearl 2000, Spirtes, Glymore, and Scheines 2000]. In this paper, we first describe the development of a theoretical model for quality of schooling using a Bayesian network approach and then discuss learning of the network structure from data using the theoretical model as a guide.

A Theoretical Model for Quality of Schooling

We use the Bayesian network modeling approach to explain learning outcomes in terms of conditions and processes within schools that are difficult to observe directly. In order to construct the model, we define the amount of learning of curriculum content attributed to schools as a function of how much time is spent by students on learning that content and the rate at which students learn. Building upon insights of Carroll [Carroll, 1963], this conceptualization makes it possible to describe causal relationships between student and teacher characteristics and behavior, students' family experiences, school and community contextual variables, and learning outcomes.

School effectiveness refers to the achievement of the system's objectives, for example learning of specified contents, skills and values. Our focus, therefore, is on factors that affect the amount of time students spend on learning curriculum content. Schools are organized to provide opportunities for learning, principally through

teaching but also through self-instructional methods. Students also can learn the curriculum outside schools, through teaching provided by others and by self-instruction.

A partial conceptualization of our model is shown in Figure 1. This network explains student learning in terms of interactions among many state variables that represent the state of affairs of the education system in a country. The effect of input or observed variables (e.g., *student attendance*, *teacher attendance*, *class size*, *teaching experience*, *family involve*, etc.) on the student learning outcome (which is measured by the output variable *reading score*) is mediated by a number of hidden variables (*teacher skill*, *motivation to learn*, *engagement attention,* etc.)

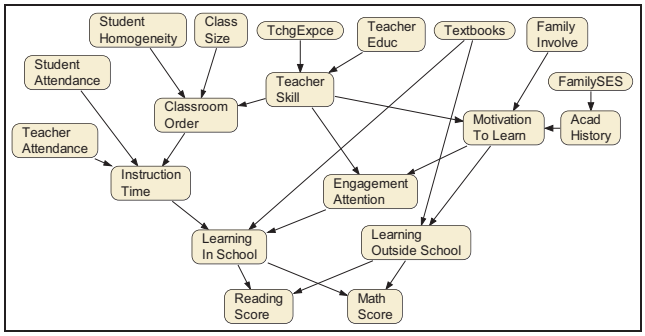

Figure 1: A partial network structure

Hidden variables (variables difficult to observe and not included in most available data) greatly reduce the number of probability estimates required to specify the network. But the main reason for introducing these variables is that they explain the causal structure of our learning model. A hidden variable such as *Instruction Time* is affected by observable variables *Teacher Attendance Student Attendance*, and *Classroom Order* but can also be affected by other interventions . For example, instruction time can be increased directly by hiring additional teachers or lengthening class time. Clearly, no general model can account for all possible interventions and include them as observable variables.

We applied the model in two case studies in Peru and Jamaica sponsored by USAID [McGinn and Moussavi 2008] to predict the reading comprehension score for 2nd grade students and to determine which interventions have the most impact on the outcome. The dataset used was collected by another USAID project [Crouch, 2008] for 512 students from 64 schools in Peru and 384 students from 48 schools in Jamaica. The model developed for these countries consisted of more than 60 variables of which only about one third were observed in the dataset. We defined all variables as binary. For example, *Textbooks* (available, not available), *Instruction Time* (adequate, not adequate), etc.

We specified the conditional probabilities according to experts' opinion, the literature review, and in a few cases based on automated learning from available data. We were encouraged that our model did as well as a regression model based on the same data set in predicting students' pass/fail reading scores. As the data sets available for Peru and Jamaica were small and very limited data was available on many variables defined in the model, we could not rely on automated learning of all the conditional probabilities from data. But the main purpose of our project was to develop a model that reflects the existing research in the field and demonstrates the impact of various interventions, both on the observed variables such as textbooks as well as on the hidden variables such as learning outside school.

Learning the Structure of the Model from Data

We have conducted a series of studies in learning the Bayesian network structure directly from available data sets using the Tetrad modeling tool [Scheines, et al., 1994]. As a start we have used data available from NELS 88 [National Center for Education Statistics, 2002] which is based on information from 11,384 American 8th grade students. The dataset includes observations for all but the following four variables shown in the theoretical model shown in Figure 1: *Teacher Skill*, *Learning outside School*, *Learning in School*, and *Instruction Time*.

We carried out a number of learning experiments both unsupervised and supervised on this dataset. Again we limited the variables to binary values. The example shown here is based on a supervised learning taking advantage of the "knowledge tiers" option provided in Tetrad. In this example we specified the following tiers:

> Tier 1: FamilySES, UrbanResidence, Teacher Attendance, TchgExpce, TeacherEduc, Student Homogeneity, ClassSize, Textbooks, AcadHist.
>
> Tier 2: Family Involve, Student Attendance, Engagement Attention, Motivation to Learn, Classroom Order.
>
> Tier 3: Read8 and Math8 (that is, the scores for reading and math).

Variables in a tier cannot influence variables in tiers above them. For example, variables in tier 3 cannot influence any other variables. We also used the option of forbidding links between variables in both tiers 1 and 3. For example, FamilySES and Teacher Attendance within tier 1 cannot influence each other.

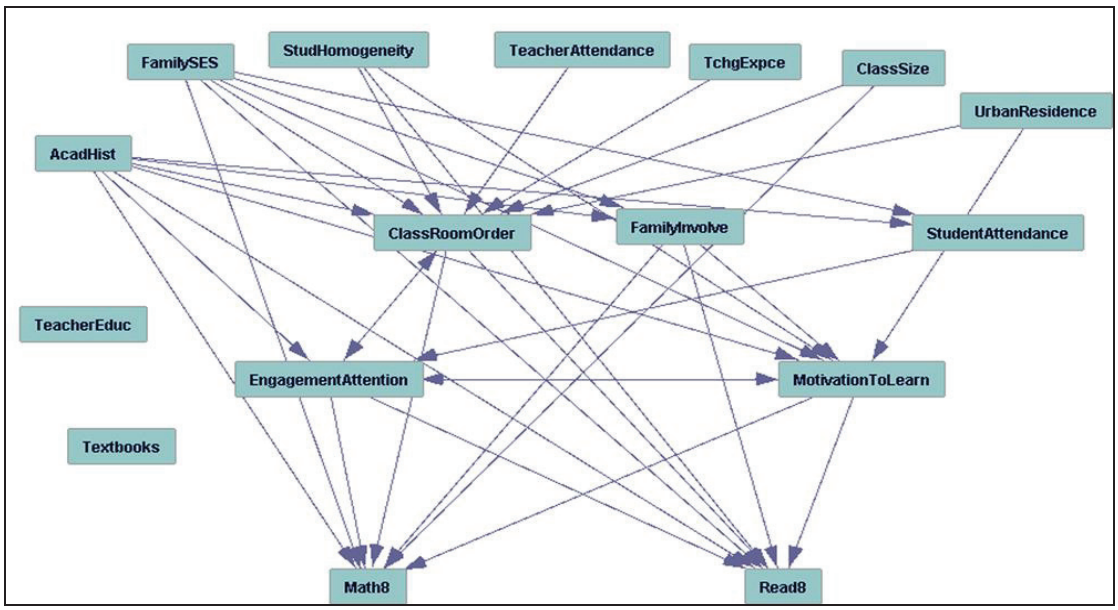

Figure 2: A discovered network structure

The resulting network structure without any latent variables is shown in Figure 2. How sensible is this result? Tetrad's discovery of the structure is based on the PC algorithm and we have not yet applied other more recent algorithms (e.g., the max-min hill climbing algorithm, greedy equivalent search) to compare the effectiveness of various algorithms. Nevertheless, it is encouraging to see that this structure is mostly in agreement with our theoretical model discussed earlier.

A number of observations can be made on the discovered structure. For example, in this network structure TeacherEduc and Textbooks seem to not matter for learning outcomes. While research has been inconsistent on the impact of teacher academic education on learning outcomes, it has shown that textbooks do matter. Irrelevance of textbooks in this data set can be explained by a number of reasons: it is possible that teachers were not using the textbooks; the contribution of textbooks to learning may depend on how (and how well) teachers use the books in their teaching; there may be another, unspecified or latent variable that accounts for both textbook use and learning outcomes; textbooks may vary in quality and in content as there are four major textbook publishers in the United States; and finally the impact of textbooks may also be linked to student use outside the classroom.

Another interesting observation is the presence of double headed links between Motivation to Learn and Engagement Attention and between Classroom Order and Engagement Attention. This indicates the possibility of a common cause for these variables. That variable could indeed be Teacher Skill as defined in our theoretical model and/or the instructional methods used by the teacher.

Finally, the presence of a link from UrbanResidence to ClassRoomOrder seems dubious and not substantiated with research in the field. However, the link from UrbanResidence to MotivationToLearn is a curious one. While to our knowledge there is no research in this area, it can be justified in the sense that schools' curricula are often more geared toward urban areas and thus more applicable to students in urban areas.

We plan to further analyze these possibilities with more extensive datasets and variables and reconcile the discovered structures with our theoretical model.

Conclusion

Many of the features of the teaching and learning process are difficult to observe and not measured in large-sample surveys of school operation and student learning. Most current policy analysis relies on data that describes only some of the material and human resource inputs to the school and characteristics of students. These factors interact in unspecified ways in the complex process of instruction and learning, and are insufficient to explain most of the variation in measures of learning outcomes. This complexity is seen in our analysis of different data sets. Schools achieve relatively equal levels of effectiveness (average student test scores) with widely differing levels of inputs and combinations of instructional

practices. Our model reflects at least some of the complexity of teaching and learning.

A model of the kind we have presented can be used in many ways. First, it can distinguish between learning attributable to a school's effectiveness, and that which is explained by learning occurring outside the school. Second, the model can be used to suggest different strategies for improving learning, some that change inputs and instructional practices in schools, others that change the school's relationship with families and the community. Third, in cases where no reliable or standardized test scores are available, the model can be used in a predictive fashion to determine identify "failing schools." Finally, the model can be used as a practical tool for educating policymakers and school administrators.

As for future work on the model, many challenging issues remain. We would like to apply more recent, state of the art algorithms to the same data set to discover the underlying network structure and compare the results against our current structure. Furthermore, we intend to conduct the analysis for a larger number of variables. In addition, we would need to validate the model and measure its accuracy in predicting the test scores.

We also need to develop dynamic models. Education by nature is a long term process and requires models that can take advantage of time series data. In this regard, of course, collection of data for the same schools over a number of years and the quality of data sets are key challenges. At present we have available, in addition to NELS 88, data from the Education Longitudinal Study of 2002, which includes data on a large sample of students over three time periods. The ELS data set includes a similar set of contextual and process variables as NELS 88. We are looking for other large data sets that include a broader range of contextual variables.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful and constructive comments.

References

Carroll, J. (1963). A model of school learning. Teachers College Record, 64, 723-733.

Crouch, L. (2008). The Snapshot of School Management Effectiveness (SSME). www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=163

Hanushek, E. A. (2008). Education production functions. In *The New Palgrave Dictionary of Economics*.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In Jordan, M. I. (Ed.), *Learning in graphical models*. Kluwer, Dordrecht, Netherlands.

McGinn, N. and Moussavi, M. (2008). Summative Report for USAID. www.causallinks.com.

National Center for Education Statistics (2002). National Education Longitudinal Study: 1988-2000 data files and electronic notebook system. Washington, DC: U.S. Department of Education.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.

Scheines, R., Spirtes, P. Glymour, C., and Meek, C. (1994). Tetrad II: User manual, Lawrence Erlbaum, Hillsdale, New Jersey.

Spirtes, Glymore, and Scheines (2000). Causation, Prediction, and Search, 2nd ed. New York, N.Y.: MIT Press.