Causal Inference Analysis of Family Educational Expectations and Students' Cognitive Abilities ——Based on CEPS Data

Xi Chen*

Department of Education Information Technology, East China Normal University, Shanghai, China

Abstract: Based on data from the China Education Panel Survey (CEPS), this paper used the SCM (structural causal model, SCM) to study the causal inferences between family educational expectations and students' cognitive abilities. The SCM model, originally proposed by Judy Perl, is effective in inferring causal relationships between variables to help policy makers develop appropriate policies and measures. We found that there is a direct causal relationship between family educational expectations and students' cognitive ability. Family characteristics variables such as: family economic level, family cultural capital, and parental education level are similarly causally related to family educational expectations and students' cognitive ability. The study result can be used to improve students' cognitive ability.

1. Introduction

In contemporary society, evidence-based policy-making increasingly relies on the acquisition of causal evidence. Causal inference methods have rapidly developed and matured over the past two decades, gradually replacing traditional correlation analysis methods and becoming the main method in the field of micro econometrics.

Causal inference analysis is statistical method for studying causal relationships, and is used to determine whether one factor directly affects another factor rather than simply observing the correlation between two factors. Causal inference analysis is a crucial method for establishing cause-and-effect relationships.

Causal analysis is important for both theoretical and practical reasons. Theoretical research tries to understand the underlying mechanisms of causal relationships, while practical research aims to identify interventions that can improve outcomes. Causal analysis can help policymakers understand the potential impact of different interventions, such as public policies, medical treatments, or marketing strategies.

In recent years, the relationship between family educational expectations and students' cognitive abilities has received extensive attention. Previous studies have shown that parental educational expectations have a significant positive impact on students' cognitive abilities[1,2], meaning that higher family educational expectations are associated with stronger cognitive abilities in students. Therefore, family educational expectations are considered one of the key factors that influence students' cognitive abilities.

However, current research also reveals some controversies in this field. Some studies suggest that the impact of family educational expectations is not unidirectional but is instead influenced by students' cognitive abilities. That is, parents tend to have higher expectations for their children when their cognitive abilities are stronger. Furthermore, some studies indicate that other factors, such as socioeconomic status, may also affect the relationship between family educational expectations and students' cognitive abilities [3]. These controversies require further investigation.

The correlation analysis between family educational expectations and students' cognitive abilities is only the basic step in exploring their complex relationship. Causal inference analysis provides a comprehensive approach to considering various factors and exploring the interplay of variables. This approach is crucial for developing effective policies and measures aimed at enhancing students' cognitive abilities and achieving effective improvements in family education.

2. Related Works

The methods of causal analysis can be divided into two types: causal analysis based on association and causal analysis based on causal inference. The causal analysis methods have been discussed in detail in domestic and foreign research, so this paper only briefly introduces the related methods.

2.1. Causal analysis based on association

Causality analysis based on association is mainly inferred by observing the correlation between variables and analyzing the relationship between variables. The

* corresponding author: 1695620896@qq.com

common methods include correlation test, OLS regression, analysis of covariance, and factor analysis.

Among them, regression analysis is one of the most commonly used causal analysis methods based on correlation, and its main idea is to explore the relationship between variables by modeling the linear relationship between one or more independent variables and the dependent variable. For the correlation analysis between family education expectations and students' cognitive level, there have been many empirical studies that prove a strong correlation between the two. Although some research show a strong correlation between the two, the two may not be causally related. Family educational expectations may simply be a proxy for other background characteristics of children, such as family socioeconomic status or parental education level, both of which are highly correlated with students' cognitive level and family educational expectations. Researchers have found that family factors such as family socioeconomic status are all related to students' cognitive levels, so an analysis is needed to explore what the relationship between family educational expectations and students' cognitive levels might be if these background characteristics are taken into account.

In fact, in terms of statistics and the direction of current research, it is inevitable that the traditional correlation analysis method is in the process of being eliminated.Traditional correlational analysis, also known as "data-based analysis," can reveal the quantitative relationships between the independent and dependent variables, but it cannot provide a reliable answer to the question of whether or not a particular factor is responsible for the formation of realistic outcomes, and thus cannot guarantee that a policy based on correlational analysis is an effective policy intervention. Policy makers who mistakenly view outcomes as causally relevant are prone to ineffective policy interventions.

The ultimate purpose of causal analysis in education is to provide certain policy interventions, but data-based analysis methods are not certain to give effective policy interventions and have certain limitations. Therefore, there is a greater need for data- and empirical-based analytical methods, namely causal inferential analysis, to help educators make better decisions.

2.2. Causal analysis based on causal inference

The application of causal inference in empirical research has been studied in the international sociological community from an early stage [4,5]. While earlier studies on the economics of education focused on the instrumental role of education in promoting individual income and socioeconomic development, in recent years, more and more scholars have begun to focus on the extent to which education is realized as an inherent human right, and to explore how public education policies should be formulated in order to effectively improve students' academic performance and the quality of teaching and learning in schools. [2] There is a growing body of economics literature on the evaluative nature of educational policies or programs, covering macro and micro areas such as market-oriented reforms in education, small class size policies, teacher training programs, and innovations in teaching methods.

Since 2010, related issues have gradually gained attention in Chinese sociological circles. For example, Yunsong Chen and Xiaoguang Fan [6] have systematically introduced the sources and solution ideas of endogeneity problems affecting causal inference. The classical methods such as instrumental variables and propensity value matching are introduced by Anning Hu [7] and thematically. Meanwhile, empirical applications of causal inference have been accumulating in the research of Chinese scholars. In recent years, domestic scholars have further explored the path of introducing new technologies such as big data and machine learning into causal inference, and the exploration in this area has kept pace with the international sociological frontier.

Among the commonly used methods, randomized controlled trials are one of the most commonly used causal analysis methods based on causal inference. The main idea is to determine the effect of an intervention on the dependent variable by randomly assigning experimental and control groups and controlling for the interference of other variables.

Propensity score matching is an assessment of the effect of an intervention in an observational study by estimating the probability of an individual being intervened on and thus finding a control individual that is most similar to the intervened individual. The propensity score matching method was first proposed by Donald Rubin, an American economist, and is mainly used to deal with missing data and imbalance between experimental and control groups.

3. Our Methods

3.1. Data and variables

The data used in this paper comes from the China Education Panel Survey (CEPS) baseline data of the 2013-2014 school year, which was provided by the China Survey and Data Center of Renmin University of China [8]. In the baseline survey, CEPS adopted a multi-stage probability proportional to size sampling method (PPS sampling) and randomly selected 438 classes in 112 schools in 28 counties across China, with a sample size of 10,279 students in grade 7 (junior high school). The CEPS uses questionnaires as the main instrument to survey all surveyed students and their parents or guardians, classroom teachers, main classroom teachers, and school officials. According to the research design, this paper selects data from the student questionnaire for analysis. This survey collects information on the basic profile, household registration and mobility, growth experience, and academic achievement of the students interviewed, as well as data related to the students' families, classes attended and schools.

3.1.1. Outcome variable

The outcome variable in this paper is the standardized cognitive ability test score (using the 3PL model) from the baseline survey. CEPS designed a set of cognitive ability test questions for 7th and 9th-grade students that don't test students' literacy knowledge, but rather measure students' logical thinking and problem-solving skills. The questions are internationally comparable and nationally-comparable. Students receive a raw total cognitive ability test score after answering the questions, and the transformed standardized score is used as the outcome variable to measure students' cognitive ability in this paper.

3.1.2. Causal variable

The causal variable in this paper is parents' educational expectations, which are scored on a ten-point scale.

3.1.3. Confounding variables

Referring to the existing research [1], other factors that may affect students' cognitive abilities are also listed as confounding variables in this paper. Since this paper focuses on the influence of family educational expectations on students' cognitive abilities, the selected confounding variables are family characteristics variables, which include:

1. Family economic level

Family economic level is the subjective judgments of students' current family economic level in the questionnaire, from 1 to 5 indicating very difficult, relatively difficult, moderate, relatively rich and very rich respectively. A rich family can broaden children's horizons and give them better resources, leading to high cognitive ability of children.

2. Family cultural capital

In this paper, the book collection is used to represent the cultural capital of students' families, and the five categories from 1 to 5 represent very little, relatively little, average, relatively much, and a lot, respectively. A high book collection implies the family has sufficient cultural capital and a strong cultural atmosphere, which may lead to high cognitive ability of children.

3. Parental education level

CEPS classifies education levels into nine categories, from 1 to 9 representing no education at all, elementary school, junior high school, junior college/technical school, vocational high school, high school, college specialist, college undergraduate, and graduate and above. In this paper, the mean of father's and mother's education level is used as the parental education level variable.Families with high levels of education may have a more academic atmosphere and parents may educate their children in a more scientific way, leading to high cognitive abilities. Families with high levels of education may receive the benefits of high education, which may lead to a greater emphasis on education, resulting in high family educational expectations. Also, such families may have a more academic atmosphere, and parents may educate their children in a more scientific way, leading to high cognitive ability of children.

3.2. The causal analysis method used in this paper

The method used in this paper is a causal analysis approach based on the structural causal model [9]. Structural causal models provide an integrated framework for causal inference, using causal diagrams to identify causal relationships from purely observational data. Causal graphs explicitly state the direction of causal associations between variables in a system and in doing so reveal non-causal associations. Originally proposed by computer scientist and Turing Award winner Judy Pearl, causal graphs have been gradually developed and refined by scholars in different fields, with the most prominent contributions from scholars related to computer science and epidemiology. In fact, we are no stranger to the logic and motivation behind causal graphs. Cause-effect diagrams visualize the association between different variables in a pictorial way through the combination of three elements: nodes, line segments and arrows, and are formally similar to path diagrams based on structural equation models.

Sociologists such as Morgan and Winship [10] were early to realize the important value of causal diagrams in understanding causal problems and worked to promote this method in the social sciences. In the domestic sociological community, Guodong Jue and Yunsong Chen systematically introduced the causal diagram method and elaborated on the concept and elements of causal diagrams.[11]. In fact, so far, causal diagrams have received little attention in the social sciences and have hardly attracted the attention of the domestic sociological community. The above causal inference framework still relies on a large number of algebraic derivations for the understanding of causal situations and the interpretation of related methods, which is not very friendly to scholars and readers studying social sciences and objectively limits the acceptance and influence of causal inferential thinking in the discipline. Realistic causal situations are often so complex that even systematically trained statisticians are inevitably difficult to express them. Therefore, this paper wants a set of intuitive and rigorous expressions of causal problems for the causal analysis of this study.

We argue that a non-parametric causal inference framework based on causal diagrams can fill the gaps of existing frameworks and provide humanities and social science researchers with a useful alternative perspective on causal inference, helping those interested in causal issues to delve into specific contexts and develop causal thinking. Therefore, this paper will use this framework for causal analysis.

3.3. Causal Inference Analysis Based on Cause-Effect Diagrams

3.3.1. Construction of cause-effect diagrams

Based on the existing studies, the preliminary construction of a causal diagram was conducted in this paper. The

following diagram depicts the causal relationships between parental education level (X1), family economic level (X2), family educational expectations (X3), family cultural capital (X4) and students' cognitive ability (X5)（like Figure 1）.

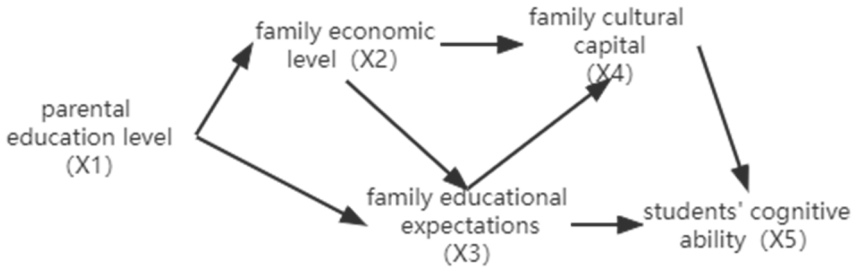

Figure 1. Cause-effect diagram based on empiricism

3.3.2. Construction of structural equations

Structural equation modeling is a standard tool in economics and sociology. In a set of structural equations, if each variable has a unique equation in which that variable appears on the left side of the equation (called the dependent variable), then the model is called structural equation modeling or causal model for short.

The corresponding structural equation is constructed based on the causal relationships in the causal diagram above and consists of five functions, each representing the autonomous mechanism controlling a variable. The variable x1, which refers to parental education level, is directly related to parental education level u1 in CEPS, and the function f2 defines the causal relationship between parental education level variable x1 and family economic level variable x2. The definition of f2 indicates that x2 will not only be influenced by x1, but may also receive the influence of error variable u2, and the meanings of f3~f5 follow in the same way.The structural equation is as following:

$$x_1 = u_1 \quad (1)$$

$$x_2 = f_2(x_1, u_2) \quad (2)$$

$$x_3 = f_3(x_1, x_2, u_3) \quad (3)$$

$$x_4 = f_4(x_3, x_2, u_4) \quad (4)$$

$$x_5 = f_5(x_3, x_4, u_5) \quad (5)$$

3.3.3. Solution of structural equations

This section describes the process of solving the structural equations defined in Section 3.3.2. The main purpose of the solution is to determine the specific forms of f2,f3,f4 and f5 in the set of equations, so as to achieve a quantitative description of the causal relationship between the variables.

Based on the real data of each variable in CEPS, the forms of equations f2,f3,f4 and f5, such as linear forms, nonlinear forms, etc., can be roughly determined. The purpose of this paper is only to explore the feasibility of causal inference techniques to be applied in the field of education. Therefore, to simplify the solution process, f2,f3,f4 and f5 are all defined as linear functions. f2,f3,f4 and f5 are defined as follows:

$$x_2 = f_2(x_1, u_2) = a_2 \cdot x_1 + u_2 \quad (6)$$

$$x_3 = f_3(x_1, x_2, u_3) = a_3' \cdot x_1 + a_3'' \cdot x_2 + u_3 \quad (7)$$

$$x_4 = f_4(x_3, x_2, u_4) = a_4' \cdot x_2 + a_4'' \cdot x_3 + u_4 \quad (8)$$

$$x_5 = f_5(x_3, x_4, u_5) = a_5' \cdot x_3 + a_5'' \cdot x_4 + u_5 \quad (9)$$

x2 is linearly related to x1 in one element, and the values of parameters a2 and u2 can be calculated directly by the least squares method based on the data corresponding to the x2 and x1 variables in CEPS. x3 is linearly related to x1 and x2 in two elements, and the values of parameters a3, a3 and u3 can be calculated based on the data corresponding to the x3, x1 and x2 variables in CEPS as well. By analogy, all parameters in the structural equations can be solved by this method. Finally, we can obtain the specific expressions for f2, f3, f4 and f5.

Based on the function expression of f5 obtained by the final solution, we can know the quantitative relationship between variables x1, x2, x3, x4 and the target variable x5. For the variables not contained in f5, it means that there is no causal relationship between this variable and the target variable x5. For the variables contained in f5, the causal relationship between them and the target variable x5 can be verified by the do() intervention operation. That is, the values of other variables are solidified first, and all observations corresponding to the variable x to be intervened are brought into the f5 expression, and if the value of x5 will not change significantly, it indicates that there is no obvious causal relationship between the variable x and the target variable x5. In both cases above, the edges between the variable and the target variable need to be removed from the causal diagram. All other causal relationships between variables can be deleted by this step.

3.3.4. The result

Based on the relevant variables and data in CEPS, the solution structure equation is solved and the final causal diagram obtained is shown in figure 2. From the graph, we can see that family educational expectations have a direct causal relationship with students' cognitive ability, while the three variables of parental education level, family economic level and family cultural capital have a direct causal relationship with family educational expectations and also with students' cognitive ability.

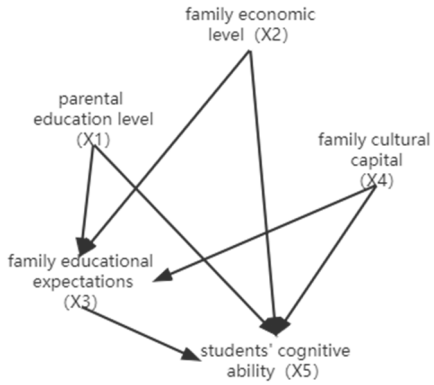

Figure 2. Cause-effect diagram based on CEPS data

4. Conclusion

This paper explored the relationship between family education expectations and students' cognitive abilities using a causal inference analysis method based on causal graphs. The main findings are: (1) Family educational expectations can directly affect students' cognitive ability, and there is a causal relationship between the two. (2) Family characteristics, including parental education level, family economic level, and family cultural capital, can directly affect students' cognitive ability.They can also indirectly affect students' cognitive ability by influencing family educational expectations. Therefore, the relevant policies based on these factors may be effective interventions for cognitive abilities.

The main innovation of this paper is that few studies have used causal inference methods based on causal maps in education, and this paper confirms the feasibility of this method for analyzing educational issues and for formulating relevant educational policies. The findings of this study can help policy makers to develop relevant interventions from the family perspective to improve family education and thus promote better cognitive development of students.

References

1. Pearce R R.Effects of Cultural and Social Structural Factors on the Achievement of White and Chinese American Students at School Transition Points[J]. American Educational Research Journal, 2006, 43 (1) :75-101.
2. Jiang Yali.Parents' Expectation, School Types and Migrant Children's Academic Achievement[J].Youth Studies,2017,No.413(02):11-18+94.
3. Gong Liyun,Wang Jingying,Peng Cong.The effect of family capital on middle school students' cognitive ability[J].Journal of Educational Development2017,No.603(02):29-33.DOI:10.16215/j.cnki.cn44-1371/g4.2017.02.006.
4. Sobel,Michael E.1996"An introduction to Causal Inference"Sociological Methods and Research 24(3):353-379.
5. Winship,Christopher and Stephen LMoran.1999.The Estimation of Causal Effects from Observational Data. Annual Review of Sociology 25:659-706
6. Chen Yunsong,Fan Xiaoguang.Measuring the Labor Market Effects of Social Capital:A literature review and research strategy of dealing with the endogeneity problem[J].Sociological Studies,2011,25(01):167-195+245.DOI:10.19934/j.cnki.shxyj.2011.01.007.
7. Hu Anning.Propensity Score Matching and Causal Inference: A methodological review[J].Sociological Studies,2012,27(01):221-242+246.DOI:10.19934/j.cnki.shxyj.2012.01.010.
8. http://ceps.ruc.edu.cn/
9. Pearl, J. 2009. Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.
10. Morgan,Stephen and Christopher Winship.2014.Counterfactual and Causal Inference New York:Cambridge University Press
11. Ju Guodong,Chen Yunsong.The Logical Power of Graphs:The Concept of Causal Graphs and Their Applications[J].Society,2022,42(03):195-221.DOI:10.15992/j.cnki.31-1123/c.2022.03.005