Böddeling, Annika; Witte, Benjamin

Working Paper

An investigation into the causal relation between institutions and economic development

Discussion Papers, No. 9/2011

Provided in Cooperation with:
Witten/Herdecke University, Faculty of Management and Economics

Suggested Citation: Böddeling, Annika; Witte, Benjamin (2011) : An investigation into the causal relation between institutions and economic development, Discussion Papers, No. 9/2011, Universität Witten/Herdecke, Fakultät für Wirtschaftswissenschaft, Witten

This Version is available at:
https://hdl.handle.net/10419/49948

discussion papers
Fakultät für Wirtschaftswissenschaft
Universität Witten/Herdecke

Neue Serie 2010 ff.
Nr. 9 / 2011

An investigation into the causal relation between institutions and economic development.

Annika Böddeling, Benjamin Witte

discussion papers
Fakultät für Wirtschaftswissenschaft
Universität Witten/Herdecke
www.uni-wh.de/wirtschaft/discussion-papers

Adresse der Verfasser:

Annika Böddeling
a.boeddeling@gmx.de

Benjamin Witte
bwitte@andrew.cmu.edu

Abstract

This paper examines causal inferences encountered in economic theorizing on the interplay of the quality of a country's institutional setup and that country's economic performance. The main focus is on one of the most influential contributions to institutional growth economics: The article "Why do some countries produce so much more output than others" by Robert E. Hall and Charles I. Jones. We will take a close look at the method applied and use Tetrad to analyze the data used for underlying causal structures. We will show that there are major weaknesses in Hall's and Jones's methodology and in the causal model they assume, place in question the causal inference Hall and Jones present in their paper.

Content

1. Introduction

The overall goal of this paper is to scrutinize the way of dealing with the matter of causation in social sciences. We realized that it is rather common to infer causation from empirical data in social sciences; and we wonder how causal inferences are commonly made in social sciences and whether or not they are valid. To approach these general questions we will examine an important contribution to growth economics: "Why do some countries produce so much more output than others" by Robert E. Hall and Charles I. Jones (1999). Applying the method of instrumental variables, Hall and Jones argue in favor of revealing institutions as a fundamental cause of long-run economic growth.

The impact of institutions on a country's wealth has not been placed in question in more recent growth economics; but Hall and Jones have been criticized for their approach of inferring just this causation by researchers such as Acemoglu et al. (2005) and Eicher and Leukert (2009). Our contribution will be to call attention to the need of dealing with the matter of causation in a more cautious and thoughtful way. We will do so by exposing major weaknesses both in the way Hall and Jones come to their conclusion and in the causal model they assume.

To follow the questions described above, the first chapter is devoted to the paper by Hall and Jones (1999). We will trace how the authors arrive at the conclusion that institutions do indeed have causal relevance for economic growth. To this end, we will describe their hypothesis, the statistical method they apply to come to reach this result, and the causal structure they state. In the second chapter we will exercise criticism. Hall's and Jones's analysis is based on certain assumptions concerning the associations between exogenous and endogenous variables, as well as on latent and observable variables. These assumptions are necessary for the statement they seek to make. We now want to argue that other associations that are no less reasonable, but whose hypotheses are less convenient, should be considered as well, if Hall's and Jones's intention of analyzing for a causal relationship is to be deemed honest. We will then comment on their data base and show some interesting findings concerning the significance of the variables they use for their study. We will then revisit the critical points we have made in the previous chapters and include them in an adapted graphical model. In the fourth part of the paper we will test the causal structure Hall and Jones argue for. We will do this by using Tetrad, a program for creating and testing causal and statistical models. Working with the same data as Hall and Jones and applying two different algorithms, the PC-Algorithm and the FCI-Algorithm, the program will produce interesting results that support our critical deliberations while falsifying not only Hall's and Jones's structural

model, but even their major hypothesis. In the last part of the paper we will emphasize our findings. We will reach the conclusion that Hall and Jones analyze and state an economic association that is definitely interesting and commendable, but that at the same time they do not pay sufficient attention to the difficulty involved in finding causal structures. They deal hastily and imprudently with the matter of causation, and thus make a strong claim but using weak instruments.

2. Dependency of institutions and economic growth by Hall and Jones (1999)

2.1 Hypothesis

The initial question Hall and Jones pose is why different countries produce different outputs per worker and therefore have different incomes. The currently existing widely accepted approaches to the question of income differences were mainly developed by Solow (1956), who constituted the neoclassical theory of growth and Romer (1994), who influence d significantly the endogenous growth theory. In these models the causes of economic growth are seen in physical capital and human capital. Hall and Jones agree with these common determinants, but by resorting to the income per capita, they observe that the analyzed determinants fail to explain the entire variation in output per worker across countries. They only explain a certain fraction of outcome differences, so that a productivity residual remains which so far lacks a plausible explanation. Hall's and Jones's objective is to specify the remaining residual by looking for further determinants that could answer the initial question of growth differences across countries.

Their hypothesis is that social infrastructure is causally relevant for the long-run economic performance of a country.[1] By social infrastructure Hall and Jones mean political and economic institutions which determine the economic environment positively. Following the authors' reasoning, a beneficial social infrastructure provides an environment which abets productive activities and provides incentives to invest and to accumulate capital. Favorable regulations and laws protect individual output and private property from diversion. Economic growth is understood by the authors as levels rather than rates of growth. This approach makes it possible to capture the long-run performance of an economy, since it is the absolute data of a growth domestic product rather than the mere rate of change that enables, say, easier comparison of different data, and moreover are more stable to historical exceptions in the data. In

[1] The central hypothesis of this paper is that the primary, fundamental determinant of a country's long-run economic performance is its social infrastructure. (See Hall and Jones (1999) p. 95)

summary, their claim is that institutions are of causal relevance for long-run economic growth. Hence, institutions as a new determinant of growth might be referred to as an appropriate explanation for the initial question of income differences across countries. To prove this claim the authors avail themselves of a statistical method that will be introduced in the next section, to then show the concrete procedure of applying the method while at the same time seeking to state causal relevance of institutions.

2.2 Applied method

2.2.1 Structural Equation Modeling (SEM)

SEM contains two interrelated models: The measurement model and the structural model, both of which must be explicitly defined by the researcher (Gefen et al. (2000)). The structural model represents the assumed structural relationship between the dependent variable, which is endogenous and either observed or latent, and the explanatory variable, which is both either endogenous or exogenous and observed or latent. It thus relies on theoretical considerations to define the causal relationship between the variables under consideration. By contrast, the measurement model incorporates the operationalization of latent variables. It defines the assumed association between the measurable instrumental or rather indicator variables and the latent variable. The following example will serve to illustrate the differentiation between a structural and a measurement model. If you want to examine the association between ebriety and the number of car-accidents, the first step is to develop the structural model which in our example would be that ebriety causes car-accidents. Since ebriety is not directly measurable, the second step is to find indicators which can stand for the non-measurable variable. These could, for example, be the blood alcohol and the ability to respond. The measurement model thus incorporates the determination of the indicators, in our case blood alcohol and the ability to respond, which makes it possible to account for the not directly measurable variable, the consumption of alcohol. Even if we do not claim to describe the development and application of SEM in every detail, there is one difference among measurement models that should be mentioned here, owing to its further bearing on some points of criticism we want to make later. One distinguishes between formative SEM models and ones that are reflective. In reflective models, the indicator variable is influenced by the latent variable that is the direction of effect going from the latent variable to the indicator variable. This implies that a change of the latent variable is followed by a change of all of its indicator variables (Bollen (1989)). In formative models it is the other way around; that is, the latent variable is influenced by the indicator variables and a change of any single indicator variable results in a change of the latent variable (Christophersen and Grape (2006)).

Hence, the example described above represents a reflective measurement model, since the indicators are influenced by the consumption of alcohol and not the other way around. But a formative measurement model is indeed given if we chose as one indicator the amount of beer consumed, as another the amount of vodka and as a third one the amount of another alcoholic beverage, because the indicators determine the latent variable, which in our example is the consumption of alcohol.

Once the measurement model and the structural model are established and the corresponding parameters are estimated, we can test how well existent data fit the model. The question is therefore to what extent the theoretical model is supported by the sample data obtained. If the data do not fit well the model has to be rejected, of course. Some researchers however support model modification as an appropriate procedure to gain valid results (Schumacker and Lomax (2004)). If the data do fit well, this shows that the correlations found in the data are in accordance with the causation predicted by the established theory-basis (Bollen (1989)). Whether or not this result obtained by applying SEM allows any statement about the actual causal structure underlying the data is probably one of the most difficult questions discussed by SEM researchers and philosophers. Accordingly, no agreement exists concerning how to answer the question; the opinions differ widely. Among the well known asserters of the opinion that it is indeed possible to make a statement about the causal structure by applying SEM are Pearl (2000) and Duncan (1975); while critical voiced Holland (1986) and Freedman (1999) among others.

2.2.2 Two Stage Least Squares

The particular SEM technique Hall and Jones apply is Two Stage Least Squares (2SLS) (Eicher and Leukert (2009)), also called the instrumental variables procedure (Oczkowski (2007)). Having provided a general overview of SEM in the previous section, we will now outline the concrete statistical procedure Hall and Jones follow.

2SLS is a statistical technique used for the analysis of structural equation models. It is an extension of the Ordinary Least Squares (OLS) method, applied in the event that one basic assumption of the OLS method is violated, which occurs if the error term of the dependent variable is correlated with its explanatory variable, since it points towards endogeneity of the explanatory variable. To illustrate the procedure of 2SLS we first consider a simple regression model:

$$Y = \alpha + \beta X + \varepsilon \qquad (1)$$

where Y is the dependent variable, X is the independent variable, α and β are estimable parameters and ε is the error term. Note that this equation represents the structural equation model that has been described above. If X and ε are correlated 2SLS needs to be applied. This technique seeks to identify another secondary variable, the so called instrumental variable, e.g. z, which has to fulfill the following two criteria:

z must not correlate with the error term ε, which is:

$$\text{cov}(z, \varepsilon) = 0 \qquad (2)$$

z has to be strongly correlated with the original explanatory variable X

$$\text{cov}(z,X) \neq 0 \qquad (3)$$

Furthermore, instrumental variables are bound to have no direct influence on the dependent variable (Bauer et al. (2009), Hall and Jones (1999)). This assumption is crucial and should be kept in mind for our forthcoming analysis. The step of detecting instrumental variables corresponds to the step of establishing a measurement model already described in the section on SEM. Given the existence of an instrumental variable, 2SLS proceeds as follows:

In the first stage a new variable $\hat{X}$ is created to substitute the original, X. This is done through the application of the OLS method, regressing X on the instrumental variable z. In the second stage, the dependent variable Y is regressed on the newly created variable $\hat{X}$. The result is the described dependency of Y on $\hat{X}$ instead of X. This enables indirect analysis of the association between Y and X despite X violating an important assumption of OLS.

The 2SLS method explained so far will now be described for Hall's and Jones's hypothesis. The structural model depicts how institutions, indicated as I, fundamentally influence long-run economic growth. Their basic measure of economic performance is the level of output per worker, indicated as $logY/L$. Formulated as a structural equation, the structural model of their hypothesis is therefore

$$\text{log}Y/L = \alpha + \beta I + \varepsilon \qquad (4)$$

where α and β are estimable parameters and ε is the error term (p. 98).

One problem that has to be addressed is endogeneity. The authors find the notion questionable that social infrastructure is an exogenous variable, i.e. a

variable that remains unexplained by the model, and claim that it much is rather determined endogenously, probably even contingent on the performance of with respect to the economy (p. 86, 99). That would mean that feedback occurs from the output per worker to the quality of institutions. This loop is problematic since it results in a correlation of *I* with the error term ε: To illustrate this, we assume that a one unit increase of *I* leads to a one unit increase of *logY/L*. If *I* itself depends on *logY/L*, this increase now leads to another increase of *I*, resulting in another increase of *logY/L*. In equation (4), this subsequent increase of *logY/L* leads to an increase of ε, since the feedback is not controlled for. The loop is formulated in the following equation

$$I = \gamma + \delta \log Y/L + X\theta + \eta \qquad (5)$$

where *I* stands for social infrastructure, *X* for a collection of other variables, and γ and δ are estimable parameters. The authors admit that the determination is very parsimonious but do not seek to describe all the determinants of social infrastructure since *I*, the real social infrastructure, is not directly measurable and thus not determinable in further detail. The correlation between the independent variable and the error term stemming from the feedback from *logY/L* to *I* signifies a violation of a strong assumption of the regression framework and thus leads to having to apply 2SLS (p. 99). But before addressing the application of 2SLS there is another aspect that has to be considered.

The authors admit that the quality of political and economic institutions is practically not measurable, and that this confronts them with a latent variable. Therefore, before applying 2SLS they have to install a measurable variable which can stand for the not measurable *I*, which is called a proxy.

The proxy used in Hall and Jones (1999) is formed as an average of two indices, the GADP and the Sachs-Warner index (p. 97ff), which will be roughly sketched here. The former is an index of government anti-diversion policies. The authors use an equally weighted average of five variables of this index: two are related to the government's protective role against private diversion, namely law and order and bureaucratic quality. Three are related to the government's possible role as a diverter itself, namely corruption, risk of expropriation and government repudiation of contracts. The latter is an index which measures the extent to which an economy is open to international trade. Openness is measured by testing how well a country satisfies the following criteria: (i) non-tariff barriers cover less than 40 percent of trade, (ii) average tariff rates are less than 40 percent, (iii) any black market premium was less than 20 percent during the 1970s and 1980s, (iv) the country is not classified as socialist, and (v) the government does not monopolize major exports. Using the average of these

indices as a proxy of the real quality of social infrastructure we receive the following modified structural equation (p. 100):

$$\log Y/L = \alpha + \beta \hat{I} + \hat{\varepsilon} \qquad (6)$$

whereas $\hat{I}$ is determined by the real social infrastructure I and the error term υ that results from the measurement error between measured and real social infrastructure. So the proxy is described by the following equation:

$$\hat{I} = \psi I + \upsilon \qquad (7)$$

The authors state ψ to be 1: hence the combination of equation (4) and equation (9) results in equation (8). After substituting the original I by the proxy $\hat{I}$, the authors come to address the problem of endogeneity by applying 2SLS. In the first step they need to identify an appropriate instrumental variable (p. 100 ff). They state that countries that were strongly influenced by Western Europe during the time from the sixteenth through nineteenth centuries were more likely to adopt favorable infrastructure; thus they see Western European influence as a determinant of social infrastructure. Since Western European influence is not directly measurable, instruments are needed which are correlated with the extent of Western European influence. As instrumental variables describing appropriately Western European influence and therefore social infrastructure Hall and Jones choose (i), the extent to which Western European languages are nowadays spoken as a mother tongue in a country, (ii), the extent to which English is nowadays spoken as a mother tongue in a country, (iii) the distance of a country to the equator and (iv), the Frankel-Romer index which, predicts the trade share of a country depending on its geographical features. We will explicitly not question the chosen instrumental variables at this point since the next chapter will be entirely devoted to this aspect. For the first stage of 2SLS, the authors use the following equation

$$\hat{I} = \hat{\gamma} + \tilde{X}\hat{\theta} + \hat{\eta} \qquad (8)$$

in which the proxy $\hat{I}$ is regressed on the instrumental variables $\tilde{X}$. The estimable parameters are modified due to the implementation of the proxy variable. At the second stage, the endogenous variable is then replaced by the predicted values from its first stage model, and the dependent variable is regressed on the newly created variables. This results in the following equation:

$$\log Y/L = \hat{\alpha} + \beta \hat{I} + \hat{\varepsilon} \qquad (9)$$

By applying the described method the authors come to the result that a difference of .01 in measured social infrastructure is associated with a difference

in output per worker of 5.14 percent (p. 103 ff). Therefore they do not refuse the structural model. They infer from this result that social infrastructure is a fundamental cause of long-run economic growth.[2]

2.3 The causal model underlying Hall's and Jones's thesis

First of all it should be mentioned that throughout their entire article Hall and Jones do not use a graphical model depicting their thesis. Altogether the authors are rather imprecise about the theory connecting the variables they use. To make explicit the way we understand the ideas developed, we created a graphical causal model. In this model we present the variables the authors use and the relationships between these variables. Since the authors remain imprecise in many cases, we are trying to interpret the presented ideas in this model, which is intended to help both the reader and the authors, gain a clear understanding of Hall's and Jones's thoughts, as well as of the way we interpret them and, in the later parts of this article, of our criticism of the underlying method and theory.

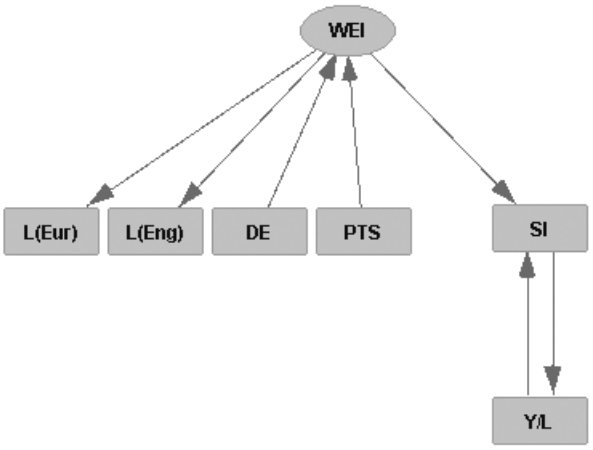

Figure 1: Graphical causal model

2.3.1 Explanation of the graphical model

The graphical model depicts Hall's and Jones's variables and their relationship as we interpreted them based on their article. As said before, such interpretation is necessary, since the authors remain rather vague. The variables used are represented by the boxes and circles (nodes). Boxes symbolize measurable variables whereas circles indicate latent variables. The arrows connecting the nodes describe the assumptions made about the causal relationships between the variables. The heads of the arrows indicate the direction of the causal relationship.

---

[2] "Differences in social infrastructure across countries cause large differences in capital accumulation, educational attainment, and productivity, and therefore large differences in income across countries" (Hall and Jones (1999), p. 114).

Starting at the bottom we have the variable "output per worker" *(Y/L)*. As a measurable variable, it is symbolized by a box, connected with two arrows: one coming from the variable "social infrastructure" (*SI*), and one leading to it. This loop describes the influence of social infrastructure on output per worker and the feedback to social infrastructure. Social infrastructure itself is determined by the latent construct "Western European influence" (*WEI*). The remaining four boxes represent the instrumental variables Hall and Jones use to measure the extent of Western European influence: "fraction of the population which speaks a Western European language as a mother tongue in a country" (*L(Eur)*), "fraction of the population which speaks English as a mother tongue in a country" (*L(Eng)*), "the distance of a country to the equator" (*DE*) and "the Frankel-Romer predicted trade-share" (*PTS*)[3].The connection between the latent construct *WEI* and the instrumental variables is not further described in the article.[4] In the following part of our paper, we will argue that it is important to define these directions and will try to establish reasonable interpretations for directions of the presented variables.

3. Point of criticism on the approach of Hall and Jones

Hall and Jones seem to be causal realists: Basic questions concerning causation do not appear in their article. Assuming as they do that causal relations exist in the world, they try to analyze one of them. Not considering the general issues of causation, it could still be worthy to challenge the causal interpretation of SEM. Though there are plenty of impassioned defenders of SEM as causal analysis, there is still open space to place the method in question as a technique enabling the detection of causal relationships. One starting point could be that even if, assuming proper application, the combination of data and structural model logically implies a certain conclusion, the causal assumptions themselves are not touched at all, but rather have to be accepted without question (Pearl (2000)). But in our opinion not the causal assumptions are what make the real statement on causation, and not the conclusion itself. It therefore appears that they are of greater importance than any conclusion consisting of a mere number indicating the magnitude of association based on already presumed causal assumptions. But tempting though it may be, this will not be the focus of our work either.

In this chapter we will focus on a different way of questioning Hall's and Jones's paper which is to examine their concrete application of 2SLS as a method of SEM. If significant flaws or even errors of application are found, general aspects

[3] See Hall and Jones (1999), p. 101f.
[4] "Our instruments are various correlates of the extent of Western European influence." (See Hall and Jones (1999), p. 100)

of causation and SEM can be disregarded and their conclusion still be placed in question, since the way they arrive at the conclusion is not legitimate.

3.1 Choice of instruments - the measurement model

One of our major concerns is Hall's and Jones's choice of instrumental variables, i.e. their attempt to build a reliable measurement model. Strictly speaking, there are three points we want to make concerning this part of their model. The first one will consider Western European influence as an adequate indicator for the emergence of institutions, the following two will have a look at instrumental variables as convenient indicators for Western European influence.

Hall's and Jones's line of reasoning for Western European influence as an appropriate indicator for the quality of institutions from a theoretical perspective is that Western Europe is the cradle of the ideas of Adam Smith, property rights, the system of checks and balances in government, in sum, of the majority of ideas nowadays deemed to be good, and in this case to be good and favorable institutions. From a statistical perspective they corroborate the chosen proxy through the correlation between Western European influence and economic institutions, which should be read as Western European influence in its role as the explanatory variable of economic institutions, since the inverse does not make sense.

One may wonder about the statement that Western European influence serves as an indicator of the quality of institutions. It could be argued that there are many countries, e.g. Togo, Bolivia or the Republic of Congo which show that high European influence does not necessarily imply the emergence of good social infrastructure. Plenty of countries that were influenced by Western European colonization have no well developed social infrastructure or high output per worker today. Western European influence as an adequate indicator seems to fit only for some countries such as the United States or Australia, where good institutions can definitely be observed nowadays. But this argument would be refuted by Hall and Jones. In their definition Western European influence has been on both the United States and Togo; but the influence of a different nature in each case. While Togo was only colonized, enslaved and exploited, the Europeans settled in the United States, bringing along their knowledge and institutions, and thus serving as the cornerstone of today's economic performance.

But there is another much stronger point to make. How can Western European influence be an appropriate indicator for Western European countries where it stands to reason that Western European influence has and always will be remarkably high in Western European countries? How can "the positive influence of a country's own historical experience upon itself" (Eicher and Leukert (2009)) be a serious indicator for the emergence of good economic

institutions? It follows logically from the causal structure hypothesized by Hall and Jones and the necessary presumption that Western European influence was and will always be high in Western European countries that the quality of institutions will always be high in Western European countries as well. This means that economic performance in Western European countries is always extraordinarily good. In addition, the extent to which explanatory variation can be found in the instruments among Western European countries is questionable. If we are right so far, the appropriateness of an indicator that by definition determines the result of a certain subgroup of the sample must definitely be placed in question. We will come back to this point later.

Even if we do accept Western European influence as a correlate of good economic institutions, the question remains as to whether the instrumental variables are well chosen. Our aim in this section is to specify the connection between the chosen instrumental variables and Western European influence, a task neglected by Hall and Jones.

The reason for taking the distance from the equator as an instrumental variable lies in the interpretation of the history of colonialism. Hall and Jones, as well as Acemoglu et al. (2005), argue that colonialists were more likely to settle in regions that were similar to the countries of their origin. They therefore preferred regions far from the equator, where geographical features such as climate more closely resembled their own. Even if examples to the contrary can certainly be found, we will not focus on arguing against this interpretation; rather, we will only state that there seems to be a connection between the distance from the equator and Western European influence. The necessity of this direction of reasoning is obvious; after all, Western European influence was not powerful enough to determine the distance to the equator of a country. Instead, the distance from the equator influenced the extent to which Europeans settled and installed their institutions, according to climate conditions similar to those in their home countries.

Another instrumental variable chosen to indicate Western European influence is the Frankel-Romer predicted trade share. Frankel and Romer (1996) investigate how international trade affects standards of living. The conclusion they draw - that trade raises income - is reached by using geographical attributes, such as countries' sizes, their distances from one another, whether they share a border, and whether they are landlocked, as instrumental variables of trade. Notwithstanding their conclusion that trade raises income, they admit that the impact is not estimated precisely. The null hypothesis that these variables have no effect is only marginally rejected. The Frankel-Romer predicted trade share is not described in further detail by Hall and Jones. In the following we will understand the index as the predicted trade share of an economy in the entire

gross domestic product, whereby the share is computed stating that it depends on the above specified geographical features.

We now want to ask why a correlation is assumed with Western European influence.[5] At first sight, a simple explanation could be that the predicted trade share is calculated from and thus contains several geographical features of a country. These features may assumed be intuitively as correlated with Western European influence, since countries such as those with a coastline were more likely to be discovered and therefore influenced by Western Europe than countries that are landlocked. But if these features are used to predict the trade share of a country which itself positively influences the economic growth of a country, the Frankel-Romer predicted trade share would be problematic as an instrument, because it might have a direct influence on the output per worker. In such case it would violate the requirement that instrumental variable do not have a direct influence on the dependent variable. But if we were to accept the predicted trade share as an instrumental variable, we could say that it is a formative instrumental variable, which would mean that it affects Western European influence and not vice versa, since it is difficult to imagine the geographical features it contains having been changed by European settlers.

There remain the instrumental variables describing languages, that is, the extent to which the primary languages of Western Europe, i.e., English, are spoken as a mother tongue. The primary Western European languages are English, French, German, Portuguese and Spanish.[6] Hall and Jones thus use two language variables. First of all we want to question precisely this use of two language variables since the latter, English as a mother tongue, is already included in the former. Does it make sense to include one and the same correlate in two different instrumental variables? Using English twice as an instrumental variable puts an emphasis on English speaking countries. If for example we have one country with English spoken by 50% of the population and another country where 50% of the people speak Spanish, the instrumental variables would indicate a higher European influence on the former country and therefore a "better" social infrastructure. That would imply that countries colonized by the British would have a higher output per worker today than, say, a country colonized by the French or the Spanish. In our opinion there is no obvious reason why that should be true, but it certainly fits the data: Countries where a significant fraction of the people speaks English today are likely to have a high output. These include America, Australia and New Zealand, all of which are highly developed countries. Spanish-speaking countries (e.g., South American countries) or French-speaking countries (e.g. Ivory Coast) are more likely to

[5] "Our instruments are various correlates of the extent of Western European influence." (See Hall and Jones (1999) p. 100)
[6] See Hall and Jones (1999) p. 100

have a lower output per worker today. But the fact that it fits the data is no reason in itself to include an instrument, and it gives the impression that the model has been modified to fit the data!

Secondly, we want to point out once more one of the problems described above: Western European languages can hardly be a good indicator for Western European influence, because in a subsample of the data set, namely the industrialized or OECD countries, they will always produce questionable results. Flawed results seem inevitable when most of industrialized countries measured are the source of the influence the instruments are supposed to measure. Used on Non-OECD countries, the instruments might have reasonable explanatory power. Differences in Western European influence may explain differences in the social infrastructure of today. But this explanatory power seems highly questionable for OECD countries, most of which are Western European.

So far we have pointed out some weak points of the chosen instrumental variables by examining their supposed correlation with Western European influence. But let us forget all the flaws of the instruments for a second and assume that the associations do indeed exist as interpreted by Hall and Jones. Although no explicit auxiliary theory that describes the relationships between the latent construct and its measures exists and we therefore lack important information from the authors (Edwards and Bagozzi (2000)), one can discover the underlying structure by analyzing the relations one by one as we have done so far. We then see that the chosen instruments have different directions of effect. This means that Hall and Jones use reflective indicators influenced by the latent construct, such as languages, along with formative indicators that determine the latent variable, such as the distance from the equator.[7]

In the literature on SEM no example or explanation can be found that legitimates the use of both reflective and formative models in one and the same measurement model. Intuitively, the combination of formative and reflective indicators seems questionable, since it could imply, for example, that the distance from the equator is indirectly responsible for the extent to which English is spoken as a mother tongue, which in turn implies a causal connection between the instrumental variables. But even if we do not consider this point any further, there is another one to make. Particularly in SEM analysis formative models can lead to serious flaws.

This results from the possibility that formative indicators may not be correlated. If we take an example such as the amount of consumed beer, wine and hard liquor as the formative indicators of mental inebriation, they may but do not need to be correlated. This causes problems with most of the commonly used

[7] See Edwards and Bagozzi (2000) p. 155 ff.

SEM techniques, since they attempt to account for the covariances of the indicators (Chin (1998)). "All items must be reflective to be consistent with the statistical algorithm that assumes that the correlations among indicators for a particular LV are caused by that LV", as Chin (1998) puts it.[8] The problem is that acceptable goodness of fit can result from SEM despite its use of formative indicators. But the validity of the results is highly questionable (Cohen et al. (1990)).

Disregarding all the problems described so far that arise from the chosen instruments there is yet another point to make. One assumption that has to be fulfilled using instrumental variables is that they do not have any direct influence on the dependent variable, in our case the output per worker; rather, they only influence it through social infrastructure (Angrist et al. (1996)). In our opinion there are interpretations of every single instrument that readily allow for violations of this assumption. We will briefly sketch them here but without going into further detail.

It is not questionable that the distance from the equator, that is to say, the geographical position, has an influence on the predominant climate of a certain region. At the same time there is a "geographical hypothesis" that claims that climate circumstances directly influence the performance of an economy by determining such things as work effort, incentives or productivity of an economy. This idea dates back to Montesquieu (Montesquieu (1989)) at least and of course, even if from a moral perspective we nowadays hesitate to believe this association, it would nevertheless show an association between the distance from the equator and the income per worker which should not exist if distance is used as an instrument. For languages it can also be argued that they indirectly influence the economic performance of a country since a greater number of native speakers of English could influence the accumulation of human capital because of easier access to knowledge and universities in developed countries or the possibility of trade both of which in turn determine the outcome per worker.

Of course, some of these lines of reasoning seem no less sophisticated than the one followed by Hall and Jones. We do not insist that they are correct or more convenient but only want to show that there are indeed reasonable interpretations that make it difficult to maintain the aforesaid assumption. These three points, the choice of Western European influence as an indicator of institutions in general, the specification of the measurement model with the described instruments, and the possibility of violating assumptions, show that the instruments chosen are rather weak and thus invite one to strongly question Hall's and Jones's way of applying the method. Moreover, there are good

[8] See Chin (1998), p. 3.

reasons to question the strong and self-confident conclusion Hall and Jones infer.

3.2 Choice of Data

3.2.1 Size of the data set

As briefly mentioned before, the data set Hall and Jones use includes 127 countries. For 79 of these countries all required data were available. The data sets for the other 48 countries lacked the value of at least one variable. These missing values were imputed from the 79 complete sets by the authors using the bootstrap method.

Throughout the literature on SEM one can find a wide range of recommended sample sizes (Schumacker and Lomax (2004)). The authors do not seem to be able to settle on a minimum satisfactory sample size. Ding et al. (1995) found numerous studies that agreed upon 100 to 150 subjects as the required sample size. "Boomsma (1982, 1983) recommended 400, and Hu, Bentler, and Kano (1992) indicated that in some cases 5000 is insufficient."[9]. Although researchers seem to be undecided about an absolute number they agree that SEM requires a much larger sample than other statistical techniques to obtain stable parameter estimates and standard errors. Hall and Jones face a common problem of growth econometrics: The number of countries in the world is rather small from an econometric point of view; and in addition to this natural restriction the required data does not exist for many countries (Durlauf et al. (2005)). Although a sample of 127, given the imputed data is valid, might be just enough, the small sample size can still lead to flawed results.

3.2.2 Style of the data set: cross-sectional data

The data Hall and Jones use are cross-sectional. In contrast to longitudinal data sets that observe one subject over time, cross-sectional data sets observe many subjects at one point in There are different implications that go along with using the one or the other kind of data. Since Hall and Jones use the former kind, we will only focus on discussing the implications and limitations arising from the use of this kind of data set.

First of all, it should be pointed out that econometric investigations typically use cross-sectional data. One of the reasons for this might be the vast amount and ready availability of these data sets. By contrast, longitudinal data often have to

[9] See Schumacker and Lomax (2004), p. 49.

be collected in studies. In addition, growth econometricians mostly try to discover the fundamental determinants of growth, which can most certainly not be found in data spanning five years.

Therefore, rather than examining the development of a country over time, Hall and Jones compare 127 different countries to find the long-run determinants of growth in the differences of their present performance. They assume that differences in the initial conditions of different countries - Western European influence - lead to their different performance today. By using instruments that in their view correlate with these differences and that "consider several centuries of world history"[10] they also assume that these differences are stable over time. That seems to be a rather difficult and questionable assumption. Instruments like the ones that analyze for languages might not be as stable over time as assumed. In Europe, for instance, the ethnolinguistic fragmentation changed completely in the era of industrialization (Acemoglu et al. (2001)). That would imply that on the one hand this variable might be endogenous since it seems to be dependent on growth. On the other hand, it clearly shows that variables are often easily taken to be stable over time although they are not.

Another general problem with the use of cross-sectional data is the assumption of parameter heterogeneity (Durlauf et al. (2005)). By interpreting the differences between many subjects the researcher automatically assumes, that these subjects are comparable. Or as Harberger puts it: "What do Thailand, the Dominican Republic, Zimbabwe, Greece, and Bolivia have in common that merits their being put in the same regression analysis?"[11] This is a general problem that the social sciences face whenever they attempt to make quantifying studies. To break it down to Hall's and Jones's paper it is questionable to make such claims as the one that a change in the extent of corruption would have the same influence on Mali as on Vietnam.

In addition, there are two further technical problems with cross-sectional data sets that we want to mention without further explanation. First, it is difficult to make inferences about the direction of causality when looking at one point in time. Knowing that the health conditions of unemployed people are worse than those of employed people does not allow us to infer that either unemployment causes bad health or vice-versa (Davies (1994)). Hall and Jones face the same problem: It is difficult to determine the direction of causation. By assuming the causation in both directions - that good institutions cause growth and growth causes good institutions - the authors avoid the pitfalls of this problem and try to analyze for this feedback using instrumental variables as described above. Secondly, cross-sectional data make it difficult to analyze for omitted

[10] See Hall and Jones (1999), p. 100.
[11] Found in Durlauf et al. (2005).

explanatory variables, since only differences between, not differences within cases are examined. As a conclusion to this discussion, we would like to point out that the analysis - although broadly used in the social sciences - is not as harmless as many researchers consider it to be.

3.2.3 A Pearson paradox?

As discussed above, it seems questionable to try to explain the output per worker in Western Europe with the extent of Western European influence. Eicher and Leukert (2009) seem to have had the same impression. In their paper they seek to analyze the difference between parameter heterogeneity OECD and Non-OECD countries with respect to the influence of institutions on growth. To do so, they take the data set Hall and Jones use and split it into OECD and Non-OECD countries. The assumption of Eicher and Leukert (2009) is that the explanatory power of the instruments will be higher in the Non-OECD countries than in the OECD countries. In their analysis they come to the not surprising result that the instruments used by Hall and Jones do not have significant explanatory value for OECD countries. To their own surprise they get the same result when testing the Non-OECD subsample. To recapitulate: Hall and Jones describe their instruments as having significant explanatory power in the complete sample of 127 countries. When split up into the two subsamples OECD and Non-OECD the instruments are not significant in either subsample. To us, this appears to be a Pearson paradox.

The Pearson paradox is named after Karl Pearson, who first discovered it in 1899. It is also referred to as the Simpson paradox since Herbert Simpson elaborated the problem in further detail. The paradox describes the phenomenon that the results of a statistical analysis may be reversed when the test group is split into subsamples (Pearl (2000) p. 78, Baumgartner and Graßhoff (2004) p. 130 ff, Aldrich (1995)). For example, a drug may be tested and the result is that it has a positive influence on the test group. But when the test group is split into subgroups, e.g. by sex, the drug shows no effect in either of the subgroups. How is that possible? The reason for this curio could be that men and women react in different ways to the drug and that there are more men or more women in the test group. So if for example women respond better to the drug and there are more women than men in the test group the result in the whole group would be a different one than the results in the two subgroups.

So when Eicher and Leukert (2009) find out that the result Hall and Jones get from their analysis - that social infrastructure is a determinant of economic growth - holds for neither subsample, we seem to have a similar situation. The instrumental variables seem to explain variation in the complete sample, but this is not the case in the subsamples. What seems to drive the result of Hall and Jones is the combination of the two groups. On the one side, there is the OECD sample. These countries are well developed and accordingly have a high output

per worker; at the same time most of these countries will also score high with the instrumental variables employed for the analysis of language and geography. On the other side there are the Non-OECD countries, whose output per worker has greater variation than that of the OECD countries but whose variation can not be explained using the installed instruments. It is clear that this problem can be seen in analogy to the Pearson paradox since the main statement is the same: the results of the entire test group are contrary to the results of the sub groups under consideration. At the same time we had difficulties in transferring the conventional structure of a Pearson Paradox, which is the contingency table. We therefore did not test to see if we are definitely dealing with a Pearson Paradox. But assuming these findings were correct, what conclusion could be drawn from them?

For the Pearson paradox example it is easy: None! It is not possible to conclude from the results whether the drug has an effect or not (Baumgartner and Graßhoff (2004)). A possible solution would be to find a common cause for both incidents. If this omitted factor could be found and added to the model feasible results could be gained. The same applies to the problem we face in the Hall and Jones analysis. One solution - and that is the one Eicher and Leukert (2009) follow - would be to reject the instruments chosen and try to look for other variables that serve better for describing the differences in economic performance. Another solution would be to question the model presented by Hall and Jones and try to find an omitted factor that drives both variables.

Although it is tempting to do so, we will not continue by exploring this apparent weakness of Hall's and Jones's work; rather, we will examine in greater detail the hypothesized causal model underlying the presented analysis. It should, however, be kept in mind that if our conjecture is correct and what Eicher and Leukert (2009) found is indeed a Pearson paradox, the results Hall and Jones present are very likely to be incorrect.

3.3 Graphical summary of some criticisms

Some of the difficulties we found while analyzing the variables used by Hall and Jones can be picked up again and included in an adapted model that thus considers some points we made in the previous chapter 2. Other points can not be included in the causal model since they only question the theoretical or content part of Hall's and Jones's analysis without suggesting other possible causal relationships.

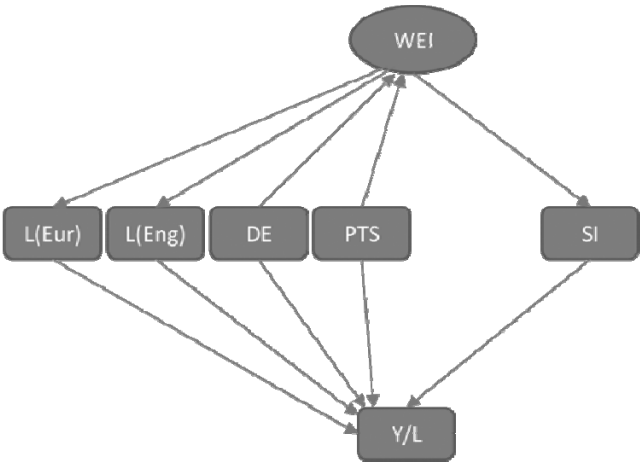

Figure 2: Adapted graphical causal model

One important difference between our model and the one Hall and Jones present is that we specified the directions of instruments and latent variables which Hall and Jones do not do. Another, even more important difference between the Hall and Jones model and ours is that we assume a direct influence of the instrumental variables on the endogenous variable "output per worker". As we have shown above, there is reason to believe that the chosen variables might influence Y/L directly. That would imply that Western European influence would not only affect the output per worker through social infrastructure, but that it could be regarded as a common cause of social infrastructure and output. To clarify our idea of how the model should be adjusted, we use two simplified, reduced models (3). The one on the left shows the causal assumptions that Hall and Jones proposed, the one on the right shows how in our opinion the assumption should be adjusted.

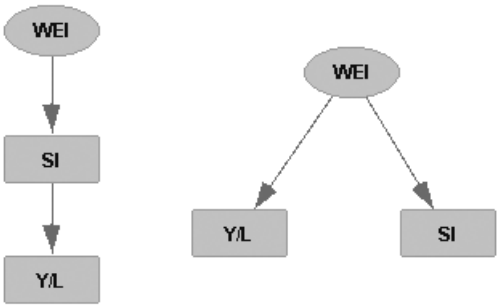

Figure 3: Reduced graphical causal models

4. SEM and causality

In this section we want to leave aside the problems we already discovered while analyzing the way Hall and Jones (1999) apply the method of 2SLS and their choice of instrumental variables. Now we are going to focus on the causal model the authors establish and the causal inference they draw from their findings.

When the first computer programs were developed to analyze causal models underlying statistical data at the beginning of the 1980s, many researchers regarded them as an easy way out of the hard business of discovering causality. Over the handiness of the programs many forgot that these were merely reducing the technical part of the work. The results were still the same as before, but quicker at hand (Cliff (1983)).

One of the basic restrictions was broadly ignored: Statistical analysis such as SEM – even if performed by a computer - can never confirm an assumed causal model, it can only fail to disconfirm it! There will always be an infinite number of models that fit the data equally well. These equivalent models will generate the same covariance matrix (Stelzl (1986)). Although statistically equivalent, only some of these models will be legitimate alternatives. Most of them can be dismissed by an intuitive understanding of causality (Cliff (1983)).

We will now try to examine whether or not the causal model supposed by Hall and Jones fits the data they used. To do so without having to calculate all the relationships manually, we will make use of the computer program Tetrad. This program was developed by Scheines, Spirtes and Glymour at Carnegie Mellon University (Scheines et al. (1998)). The current version, which is also the version we used for our test, is Tetrad IV.[12]

Besides many other functions, the program searches statistical data, whether tabular or covariance data, for DAGs. Although the user can choose from a variety of different search algorithms, in a first step we only use the classic PC-algorithm. We will then find out that the PC-Algorithm is a very restrictive one and that hence the results can only be analyzed in part and with difficulty. To address this problem we will then apply the FCI-Algorithm, which has fewer restrictions, but of course also less explanatory power.

4.1 The PC-Algorithm

One chapter in the big book of causality is titled "Probabilistic Causality". Probabilistic causality combines statistical methods with the philosophical idea of causation among variables. It assumes that every set of independences is a representation of a "directed acyclic graph (DAG)". A DAG is a construct of variables and arrows. The arrows resemble statistical relations on the one hand and causal structures on the other. The theory of probabilistic causation supplies us with tools to obtain the set of independences from a given DAG and the inverse direction (Scheines (1997)).

[12] Tetrad IV is a freeware program that can be downloaded on the Tetrad-project homepage at: http://www.phil.cmu.edu/projects/Tetrad/

With the data from Hall and Jones we have a set of independences. Now we can use the PC-algorithm to discover the possible DAGs that underlie this data set. The PC-algorithm was developed by and named after Peter Spirtes and Clark Glymour at Carnegie Mellon University. Using probability distribution with no omitted common causes as a given, it searches for statistically equivalent causal graphs (Baumgartner and Graßhoff (2004)). To do so, it makes use of two conditions: the Causal Markov Condition and the Faithfulness-Condition.

The Causal Markov Condition demands that every variable X in a set of factors V generated by a DAG G is independent of every other variable conditional on all of its effects (Scheines (1997)). Considering the set of variables X_1, X_2, X_3:

$$\{[X]_1 \perp X_3 | X_2\} \quad (10)$$

equation (7) would be mean that X_2 blocks X_1 from X_3. This implies that there is no direct causal relationship between the two. If this is found in the analysis of X_1, X_2 and X_3 there are three possible DAGs that could underlie this independence:

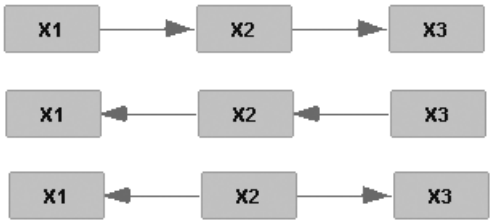

Figure 4: Possible equivalent DAGs

The Faithfulness-Condition says that every probability distribution generated by the causal structure G contains only those conditional independence relations that G contains. The assumptions are that every probability distribution generated by a causal structure fulfills the Causal Markov Condition and that every probability distribution analyzed fulfills the Faithfulness-Condition (Baumgartner and Graßhoff (2004)). The way the PC Algorithm works is that it starts with a complete undirected graph. That is a graph that includes all variables and where every variable is connected to every other variable with an undirected edge. Now the algorithm analyses every possible relation between the given variables. First of all, every edge between two variables is removed when they are statistically independent. So if for X_4 and X_5

$$P(X_4|X_5) = P(X_4) \quad (11)$$

is the case, the edge between them will be removed. After that, the remaining edges will be checked and directed by applying the Causal Markov Condition. Not every edge can be directed. But what does it mean if an undirected, directed or no edge remains? The interpretation of "no edge" is the easiest: It means that the variables are statistically and causally independent. An undirected edge, e.g. between X_1 and X_2, fits several models:

- X_1 could be the direct or indirect cause for X_2
- X_2 could be the direct or indirect cause for X_1

A directed edge from X1 to X2 is more explicit and should be interpreted as follows:

- X_1 could be the direct or indirect cause for X_2.

After the application of the PC-algorithm, some edges of the complete, undirected graph should have been removed or directed. Although there are still different equivalent interpretations of the graph, the amount is reduced and several models that would still be possible can be eliminated as non-legitimate.

4.1.1 Testing Hall's and Jones's causal model using PC

To test the causal model the authors used, we will first analyze, which relations exist between the variables in the hypothesized model and which variables block others. Our next step will be to feed Tetrad the data the authors used for their analysis. Subsequently, we will compare our results from the first step with the results Tetrad provides us and take a closer look at some relations between variables.

If we look at the causal model Hall and Jones imply, there seem to be three important connections to analyze: First of all, the link between output per worker (Y/L) and social infrastructure (SI), which is the core of the article. The question is whether or not the result will support their thesis of SI as a cause of Y/L. But two other connections are very important as well and have a major influence on the viability of the applied method. These are on the one hand the relation between SI and the instrumental variables (IV). According to the implied causal model, SI should be independent of the instruments conditional on WEI.

$$[SI \perp IV | WEI] \quad \textbf{Fehler! Verweisquelle konnte nicht gefunden werden.} \qquad (12)$$

If (9) is not the case that would imply that the instrumental variables are not chosen correctly. The conditional independence described in (9) guarantees that the feedback from Y/L to SI is controlled for and the endogeneity of SI is

"fixed". Unfortunately, the PC-algorithm is not able to control for hidden common causes, which in Hall's and Jones's causal model is Western European influence. We will test this relation at a later point.

The other relationship that is of great importance is the connection between the IVs and Y/L since this connection is a major assumption that needs to be made when applying 2SLS. The PC-algorithm is able to test this relation, and so it makes sense to apply this algorithm even if it does not depict the previous relation. Applied to all variables included, the result of the PC-algorithm looks like this:

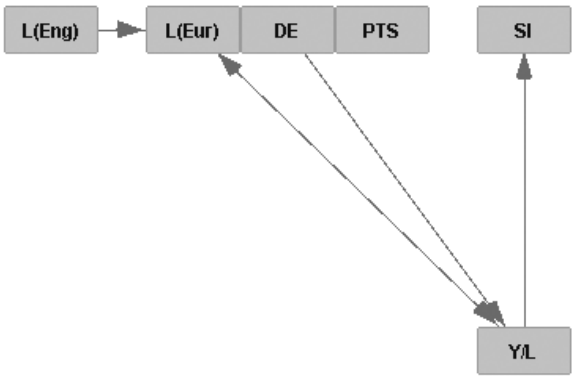

Figure 5: Results of the PC-algorithm

The edge between SI and Y/L will be discussed at a later point. First of all, we would like to focus on the edges between SI and the IVs. As one can see, there are none. In a normal case that would imply that these variables are statistically and causally independent. At this point, though, one should keep in mind that there might be a latent common cause (WEI) for which PC is not able to account. To deal with this restriction we will at a later point use a different algorithm to analyze this relationship in greater detail. Probably the most stunning result is the edges between L(Eur) and Y/L respectively DE and Y/L. In the case of DE, the PC algorithm deduces that DE is a direct cause of Y/L, i.e., that the causal effect goes from DE to Y/L with no intermediation by any other observed variable. At the same time one essential assumption of the method of instrumental variables also stated by Hall and Jones is that the instruments are correlated but causally independent and have no influence on the dependent variable, in our case Y/L. But as we argued before, the result of the PC-algorithm implies a causal relation between distance from the equator and output per worker and thus makes DE useless as an instrumental variable.

Note that PTS seems to be independent of all other variables. We will come back to this finding later in our analysis.

The bi-directed edge between L(Eur) and Y/L is the result of a partial failure of the PC-algorithm, which can be solved by either collecting more data or

introducing prior knowledge. Thus we will not focus on this aspect in more detail (Glymour et al. (2004)). But since the PC-algorithm is subject to such strong restrictions - e.g. the assumption that there is no latent common cause - we will use another algorithm on the same data set to compare the results.

4.2 The FCI-algorithm

The FCI-algorithm (Fast Causal Inference) is not as restrictive as the PC-algorithm. Its advantage for us is that its results are stable even when unrecorded (hidden, latent) common causes are involved. But every advantage also brings a disadvantage: The results of the FCI are not as well defined as the ones PC provides. Rather than DAGs it gives us PAGs (Parental Acyclic Graphs), which have to be interpreted differently (Glymour et al. (2004)). PAGs are to be interpreted as follows:

- An edge from X to Y indicates that X is a direct or indirect cause of Y
- An edge from X to Y with an arrowhead pointing to Y indicates that Y is not a cause and not an ancestor of X
- An edge with two arrowheads between X and Y indicates a hidden common cause for X and Y
- An edge marked with a dot indicates that the algorithm was unable to decide whether there should be an arrowhead or not.

Since some edges are not as defined as in the results of the PC-algorithm, a greater number of possible equivalent models result from the findings of the FCI-algorithm.

4.2.1 Testing Hall's and Jones's causal model using FCI

The FCI-algorithm applied to our data set gives us the following result:

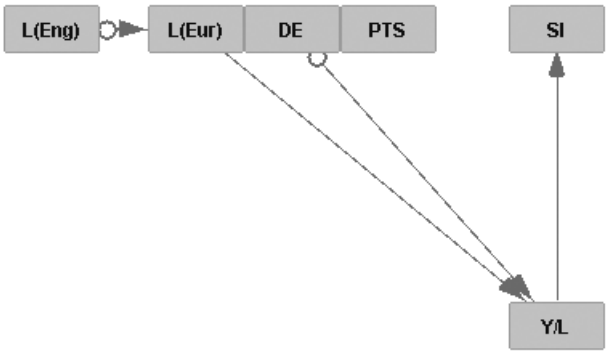

Figure 6: Results of the FCI-algorithm

The graph appears to be similar to the one found by the PC-algorithm. But this time the interpretation differs somewhat, due either to new signs that did not

appear in the application of PC, or to another interpretation of one and the same sign arising from different assumptions made by FCI. Let us start by considering the different relations going from the left to the right of figure 6.

As shown in figure 6, L(Eng) and L(Eur) are connected by an edge that is marked with a dot. As indicated before, this means that the algorithm was not able to decide whether there should be an arrowhead or not, i.e. the algorithm does not see any reason that argues against either an edge with two arrowheads or an edge with one.

If we assume an edge with one arrowhead this would mean that L(Eng) is a direct or indirect cause of L(Eur). Does this make sense? It is of course logically correct that if more people speak English more people speak a European language as well; however, this is not the case because English is the historical cause of European languages, but rather simply because English is a subset of European languages and thus correlated with them.

These content-related concerns enable us to exclude the possibility of an edge with one arrowhead thus we know that it must be an edge with two arrowheads. As previously explained, an edge with two arrowheads indicates a hidden common cause between L(Eng) and L(Eur). It is of course possible that this common cause is WEI, which would match with our considerations concerning languages as reflective indicators. The common cause implies that L(Eng) is independent of L(Eur) conditional on the common cause, which makes much more sense than the possibility we considered before. Knowing that L(Eur) is a direct cause of Y/L and that there is no direct connection indicated between L(Eng) and Y/L, we can furthermore state that L(Eng) is independent of Y/L conditional on the common cause. But of course this does not imply probabilistic independency since the probability of L(Eng) changes if we have the additional information Y/L and knowing that this implies that L(Eng) is caused by the common cause of L(Eng) and L(Eur), which then caused Y/L.

Nevertheless, one important question remains open: why is there no direct influence of L(Eng) on Y/L? How is it possible that L(Eur) causes Y/L but not L(Eng), which is part of L(Eur)? We feel unable to interpret this result and can only suppose that the reason lies in the meaningless choice of L(Eng) as an extra instrument without having to take into account that it is already included in L(Eur).

Again, we find an edge between DE and Y/L, and again the arrowhead points towards Y/L. An edge with arrowheads on both sides would mean that there is an unrecorded common cause for the two variables. Again, this does not seem to make a lot of sense. But our interpretation of DE as a cause of Y/L still fits the result.

As in the PC result, PTS is not related to any of the other variables. But is this surprising? On the one hand yes, since it includes geographical features, as does DE and might therefore have an influence on Y/L. Moreover, it should be related to SI, since, according to Hall's and Jones's model, PTS is a cause of WEI, which is a latent cause of SI. That would make PTS an indirect cause of SI, to be indicated by a directed edge. This missing edge lets us doubt the correct choice of PTS as an IV.

On the other hand, it is not surprising that this IV has no connection to the other IVs. Being as it is not a reflective instrumental variable, it does not have the common cause WEI, as the two language indices have. The edge between these two seems more plausible than in the PC-result, because it also includes a latent common cause, which we have in WEI.

It is remarkable that the two major points of chapter three - the direction of the instrumental variables and the possibility of a direct influence of the instruments on the outcome per worker - are not disconfirmed by this algorithm. The arrow between L(Eng) and L(Eur) includes the possibility of a common cause, which we argued for and which could be Western European influence. DE and PTS, however, do not have any relation, which our model also states, since our argument states that while they are possibly not determined by WEI, they do themselves have an impact on it.

Even more stunning is the result that none of the instrumental variables are related in any way to social infrastructure, which obviously should be the case following Hall's and Jones's model. Considering all the far reaching consequences of this result, we want to be cautious with its interpretation. But if we take the results seriously, they place the appropriateness of the chosen instruments much more strongly in question than we have done so far.

Since we did not interpret their edge in detail in the PC-result, we will do that now: Again, SI and Y/L are connected by a directed edge. In the PC result this meant that Y/L was the direct cause of SI, thus the opposite result of Hall's and Jones's hypothesis. Now we find the same edge, but this time with the looser interpretation that there could also be a hidden common cause, but still without supporting the authors' hypothesis. That would imply that Hall's and Jones's hypothesis is not supported by the data. However, our reduced model sketched in the graphic is not rejected by these results since the arrow from Y/L to SI includes the possibility of a common cause, which in our graphic was Western European influence. We seek to make clear though that Western European influence can only serve as a common cause either of languages or of social infrastructure and output per worker. Otherwise there would have had to be additional relations, such as those between languages and the output per worker.

What still needs to be considered is the feedback from Y/L to SI. This could be the driver of the results, since it is not accounted for by the algorithms. It means we ought not to over-interpret the result for this relation proposed by the algorithms. What conclusion do these results finally permit? Following the application of the two algorithms, we found out primarily that the data do not support the structural model, i.e., the causal model Hall and Jones use as the basis for their paper. Thus, both algorithms demonstrate a different causal structure than the one proposed by Hall and Jones. Therefore, given the data under the application of the algorithms, their thesis that social infrastructure causes long-run economic growth, is not supported. As already mentioned, one should nevertheless take care not to put too much weight on this result.

To finalize our critique, we want to sum it up in a last graphical causal model:

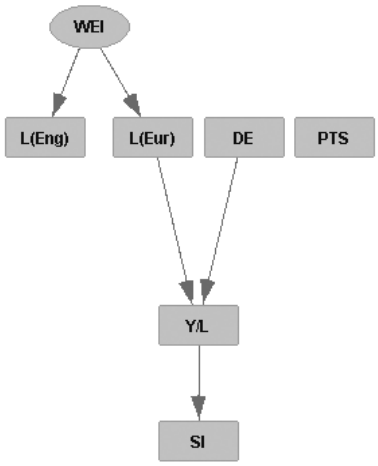

Figure 7: Combination of content and the causal results

The model displays a combination of our critique and the results of the algorithms. The common cause of the two language indices proposed by the FCI-algorithm could be Western European influence. The directed edges from L(Eur) and DE to Y/L still support our hypothesis of their direct influence on output per worker and therefore render them as useless instrumental variables.

It does not seem to be the case that the two geographical indices (DE and PTS) are the causes of WEI. If they were, the result of the FCI-algorithm would have to have edges between the geographical and the linguistic indices, since then they would be indirect causes of them. This fact and the fact that social infrastructure seems to be independent of WEI conditional on Y/L shows that the latent construct Western European influence does not have the causal relations the authors assume: If WEI is to be the common cause of (at least some

of) the instrumental variables, there would have to be an edge with two arrowheads connecting SI and the instrumental variables.

Although not all our points of critique are supported, the result of the algorithm and our hypothesis correspond in the aspect that the model Hall and Jones suppose is not supported in any way by the data. Again we want to point out that it is not our aim to question the economic theory behind the statistical analysis of Hall and Jones. We call into question the way the authors try to establish a causal relationship between the variables they observe. In our opinion, which is supported by our findings, the variables are not chosen carefully and the result the applied method proposes is over-interpreted.

5. Conclusion

The motivation to write this paper arose from an intuitive skepticism regarding the great number of researchers who deal with the matter of causation in economics. We then wanted to analyze one example in more detail to see whether or not such doubts are founded. We selected the example of Hall and Jones because their way of dealing with the issue of causation provoked particularly strong doubts while working with the paper in another context. Finalizing this paper, we see now that our initial doubts and questions were more than legitimate.

After having described the hypothesis Hall and Jones seek to account for and the method they apply, we began our content-related critique, in which we pointed out two major considerations. The first finding is that Hall and Jones use formative instruments without notice and on top of that combine them in one and the same model with reflective indicators, a procedure that is known to have serious ramifications on the validity of the measurement model. We believe there is a serious possibility that the assumption of direct dependency between the instrumental variables and the dependent variable is violated. In addition to these aspects we showed that the validity of the data set is highly questionable not only because of its size, but also because making use of it leads to results that supposedly indicate a Pearson Paradox. These major concerns stemming solely from content related deliberations show that Hall and Jones use indeed alarmingly weak instruments that are certainly not strong enough - leaving the general aspects involving the inference of causation from statistical data unconsidered - to corroborate the universally valid causal claim they want to make.

In the second part of the paper, devoted to the formal analysis of the causal structure, we left aside the previously elaborated difficulties for the sake of testing the causal model underlying their claims, using two different algorithms to do so, namely the PC-Algorithm and the FCI-algorithm. Although not all

causal relations could be considered using the PC-algorithm due to its restrictive assumptions, it was nevertheless helpful to make use of it in order to endorse the results received by using the FCI-algorithm. The findings not only confirmed our doubts on the causal model; they even intensified them. We hoped to find out that the causal model underlying the data would permit other interpretations than the one Hall and Jones hold, such as the one we suggested in the previous chapter. But when the computer program was fed with the data Hall and Jones used, both algorithms depict a causal model that is incompatible with the one Hall and Jones argue for. Thus the stunning result is thus that following the findings of the algorithms, and above all given that the algorithms are adequate, the causal model Hall and Jones state does not match the data.

Our goal was not to put into question their hypothesis itself since it seems quite plausible that an economic environment consisting of institutions that abet and enhance economic actions has an impact on the long-run economic growth of an economy. Our concern was to test the method they used to infer causation from mere data, and to examine the way they do so. We started our paper with the question whether it is legitimate to talk about a causal relationship applying the described method and we concluded with the finding that Hall's and Jones's use of the method is flawed and impaired by unconsidered ramifications, possible violations of assumptions, a statistical paradox and a disconfirmed causal structure. Thus we accuse the authors of an overhasty and rash way of stating causation. Again, as stated before, the main contribution of this paper is not the falsification of the Hall and Jones hypothesis, but rather an elaboration on the methodological shortcomings of the paper. We hope we have done this adequately and that we were able to show that it is not as easy as it seems at first sight to use statistical analyses to infer causation in social science.

Nevertheless, plenty of aspects are worthy of consideration in further detail. One aspect could be to test our conjecture that the data imply a Pearson Paradox. We only commented on the astonishing differences between the significance of the instruments in the entire sample and the subgroups without any further investigation. It would surely be an interesting aspect to challenge, as confirming that it is indeed a Pearson Paradox would significantly strengthen our argument.

Moreover, we did not investigate any further how an algorithm such as the FCI-algorithm deals with feedback. From a statistical perspective it is generally accepted that it is possible to analyze for feedback e.g. by applying the procedure of instrumental variables, as Hall and Jones do. But does an algorithm accommodate this procedure? If it does, why did the algorithms we used lead to result contrary to reached by Hall and Jones? And if it does not, how does an algorithm deal with the problem of causality in both ways? These questions

were only touched superficially but definitely need to be taken up again, since in the end the reliability of our results depends on their answer.

Besides these rather precise aspects there is of course a much broader question that needs to be investigated. We wonder whether the objective driving economics in the past decades, which is to quantify as many observations as possible and to recognize observations based mostly on such quantification, is an appropriate one. The objective of quantification presumes that social sciences work more or less identically to natural sciences, meaning that social interactions follow principles similar to physical ones and can thus be captured and depicted using the same methods. Bearing in mind all the difficulties, inaccuracies, and even errors Hall and Jones have to accept just to be able to deliver a mere number for an association that is intuitively more than obvious, we doubt whether the method they use and the goal they pursue are the right ones. To investigate this question developing a more adequate method for economics could thus be the subject of a further paper.

6. References

Acemoglu, D., S. Johnson, and J. Robinson (2001). The colonial origins of comparative development: an empirical investigation. American Economic Review, 1369 – 1401.

Acemoglu, D., S. Johnson, and J. Robinson (2005). Institutions as a fundamental cause of development. Handbook of Economic Growth, 386 – 472.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. Statistical Science, 364 – 376.

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91(434).

Bauer, T., M. Fertig, and C. Schmidt (2009). Empirische Wirtschaftsforschung: Eine Einführung. Springer.

Baumgartner, M. and G. Graßhoff (2004). Kausalität und kausales Schliessen: eine Einführung mit interaktiven Übungen. Bern Studies in the History and Philosophy of Science.

Bollen, K. (1989). Structural equations with latent variables. Chin, W. (1998). Issues and opinion on structural equation modeling. Management Information Systems Quarterly 22(1), 7 – 16.

Christophersen, T. and C. Grape (2006). Die Erfassung latenter Konstrukte mit Hilfe formativer und reflektiver Messmodelle. Methodik der empirischen Forschung, 115 – 132.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. Multivariate Behavioral Research 18(1), 115 – 126.

Cohen, P., J. Cohen, J. Teresi, M. Marchi, and C. Velez (1990). Problems in the measurement of latent variables in structural equations causal models. Applied Psychological Measurement 14(2), 183.

Davies, R. (1994). From cross-sectional to longitudinal analysis. Analyzing social and political change: a casebook of methods. London: Sage, 20 – 40.

Ding, L., W. Velicer, and L. Harlow (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. Structural Equation Modeling 2(2), 119 – 143.

Duncan, O. (1975). Introduction to structural equation models. Academic Pr.

Durlauf, S., P. Johnson, and J. Temple (2005). Growth econometrics. Handbook of economic growth 1(Part 1), 555 – 677.

Edwards, J. and R. Bagozzi (2000). On the nature and direction of relationships between constructs and measures. Psychological Methods 5(2), 155 – 174.

Eicher, T. and A. Leukert (2009). Institutions and economic performance: endogeneity and parameter heterogeneity. Journal of Money, Credit and Banking 41(1), 197 – 219.

Frankel, J. and D. Romer (1996). Trade and growth: An empirical investigation.

Freedman, D. (1999). From association to causation: some remarks on the history of statistics. Statistical Science, 243 – 258.

Gefen, D., D. Straub, and M. Boudreau (2000). Structural equation modeling and regression: Guidelines for research practice. Structural Equation Modeling 4(7).

Glymour, C., R. Scheines, P. Spirtes, and J. Ramsey (2004). TETRAD IV manual.

Hall, R. and C. Jones (1999). Why Do Some Countries Produce So Much More Output Per Worker Than Others?. Quarterly Journal of Economics 114(1), 83 – 116.

Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association 81(396), 945 – 960.

Montesquieu, C. d. S. (1748 (1989)). The spirit of the laws. Cambridge University Press Cambridge, MA.

Oczkowski, E. (2007). Two-Stage Least Squares (2SLS) and Structural Equation Models (SEM). URL http://csusap.csu.edu.au/eoczkows/home.htm, last accessed 2nd Dec 2009.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge Univ Pr.

Romer, P. (1994). The origins of endogenous growth. The Journal of Economic Perspectives,3 – 22.

Scheines, R. (1997). An introduction to causal inference. Causality in crisis, 185 – 99.

Scheines, R., P. Spirtes, C. Glymour, C. Meek, and T. Richardson (1998). The TETRAD project: Constraint based aids to causal model specification. Multivariate Behavioral Research 33(1), 65 – 117.

Schumacker, R. and R. Lomax (2004). A Beginner's Guide to Structural Equation Modeling. Lawrence Erlbaum.

Solow, R. (1956). A contribution to the theory of economic growth. The Quarterly Journal of Economics, 65-94.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. Multivariate Behavioral Research 21(3), 309 – 331.