Grossmann, Volker; Osikominu, Aderonke

Working Paper

Let the Data Speak? On the Importance of Theory-Based Instrumental Variable Estimations

CESifo Working Paper, No. 7469

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Grossmann, Volker; Osikominu, Aderonke (2019) : Let the Data Speak? On the Importance of Theory-Based Instrumental Variable Estimations, CESifo Working Paper, No. 7469, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/198829

CESifo WORKING PAPERS

7469
2019

January 2019

Let the Data Speak? On the Importance of Theory-Based Instrumental Variable Estimations

Volker Grossmann, Aderonke Osikominu

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7469*
Category 12: Empirical and Theoretical Methods

Let the Data Speak? On the Importance of Theory-Based Instrumental Variable Estimations

Abstract

In absence of randomized controlled experiments, identification is often aimed via instrumental variable (IV) strategies, typically two-stage least squares estimations. According to Bayes' rule, however, under a low ex ante probability that a hypothesis is true (e.g. that an excluded instrument is partially correlated with an endogenous regressor), the interpretation of the estimation results may be fundamentally flawed. This paper argues that rigorous theoretical reasoning is key to design credible identification strategies, aforemost finding candidates for valid instruments. We discuss prominent IV analyses from the macro-development literature to illustrate the potential benefit of structurally derived IV approaches.

JEL-Codes: C100, C360, O110.

Keywords: Bayes' rule, economic development, identification, instrumental variable estimation, macroeconomic theory.

*Volker Grossmann**
University of Fribourg
Department of Economics
Bd. de Pérolles 90
Switzerland – 1700 Fribourg
volker.grossmann@unifr.ch

Aderonke Osikominu
University of Hohenheim
Department of Economics (520B)
Germany – 70593 Stuttgart
a.osikominu@uni-hohenheim.de

*corresponding author

January 13, 2019

1 Introduction

In many fields of the social sciences, randomized controlled experiments are rare and difficult. This is particularly true for macroeconomic hypotheses that need to be tested with aggregate data.[1] Examples include the potentially fundamental role of institutional or cultural factors for economic development and international trade (e.g. Acemoglu, Johnson and Robinson, 2001; Barro and McCleary, 2003, 2006; Guiso, Sapienza and Zingales, 2006, 2009; Tabellini, 2008, 2010; Becker and Woessmann, 2009, Hanushek and Woessmann, 2012; among many others). Identification of causal effects in this and the related literature is often aimed via instrumental variable (IV) strategies, typically based on two-stage least squares (2SLS) estimations. Existence of causal effects of interest is typically viewed as being well supported by the estimation results, if *(i)* in the reduced form regression of the treatment (first stage) the coefficients on the excluded instruments are statistically significant, *(ii)* in the structural equation the coefficient on the treatment is statistically significant, and *(iii)* intuitive reasoning suggests the excluded instruments are uncorrelated with the error term in the structural equation.

This paper argues that these criteria are insufficient to gain confidence that a hypothesized causal relationship actually exists. In the language of statistical research, hypothesis testing based on these criteria alone may generate too many "false positives". In medical and pharmaceutical research, the attitude of "letting the data speak" has been heavily criticized in view of the many at first glance promising experimental outcomes of new treatments which often times could not be replicated in follow-up studies (Ioannidis, 2005). We argue that a potentially large body of research in empirical macroeconomics, particularly but not exclusively when based on aggregate data, is potentially even more problematic. According to Bayes' rule, under a low *ex ante* probability that a hypothesis at some stage of the estimation is true, the interpretation of IV results may be fundamentally flawed. That is, there may still be a high probability that the hypothesized structural relation does not exist, even when the estimated coefficients of

[1] For an exception in the context of central bank policy and money illusion, see e.g. the experimental designs by Fehr and Tyran (2001, 2008).

interest are statistically significantly different from zero. To see this, consider the null hypothesis that a causal effect from A to B does not exist and therefore that the true coefficient of interest in a regression is zero. The level of significance in, say, a t-test, is the maximum tolerated probability to reject the null hypothesis given that A in fact does not have an effect on B. What we are interested in, however, is the probability that the null hypothesis is indeed false *given that we reject it*. The chosen level of significance in a t-test is one determinant of this probability, but not the only one. It interacts with the *ex ante* probability that A causally affects B. This is the ultimate reason why letting the data speak alone can never suffice and rigorous theoretical reasoning is indispensable for deriving proper identification strategies.

Specifically, we argue that in IV analysis the credibility about a causal effect of interest is only as credible as the weakest theoretical argument among those motivating the structural equation *and* the first stage relationship. Particularly the choice of instruments in many applications is often based on simple intuitive reasoning rather than on rigorous analysis to support logical consistency of the hypothesized causal effect. We illustrate how formal theory can motivate both that the instrument is relevant and that the exclusion restriction holds.

In the coming section, we demonstrate the importance of the *ex ante* probability that a hypothesized causal relationship exists for economic conclusions by applying Bayes' rule. In section 3, we clarify the importance of Bayes' rule for instrument relevance and discuss the difficulty of finding exgenous instruments. Section 4 illustrates how to use rigorous economic theory to justify both instrument relevance and validity of the exclusion restriction in an IV approach that potentially identifies the causal effect of changing the economy's human capital stock on (long-run) per capita income. We relate our illustrative dynamic general equilibrium model to the prominent studies by Tabellini (2010) and Hanushek and Woessmann (2012) and discuss how it supports their IV strategies and results.

Our paper is part of an ongoing debate about potential flaws in making causal inferences using IV approaches. Our contribution is to shift the focus to the *ex ante*

probability that a causal relationship exists, motivated by Ioannidis (2005), and the important role of economic theory for empirical research in social sciences. Murray (2006), Angrist and Pischke (2009), and Wooldridge (2010), for instance, discuss in an accessible way the potential econometric inconsistencies of IV estimates generated by endogenous or weak instruments.[2] More specifically, Staiger and Stock (1997), Stock, Wright and Yogo (2002) and Stock and Yogo (2005) study the sampling distribution of IV estimators under weak instruments. Kiviet and Niemczyk (2014) examine the sampling distribution of IV estimators under endogeneity of some of the instruments. They conclude that instrument weakness has much stronger effects on the finite sample distribution than instrument endogeneity. When the instrument is endogenous but not very weak the finite sample distribution of the IV estimator tends to be close to normal with probability mass centered around its probability limit. Other research examines ways to test for regressor endogeneity and examines the finite sample performance of such tests, see e.g. Kiviet (2017) and Kiviet and Pleus (2017).[3]

These strands of the literature do not address, however, the potential of rigorously formulated economic theory to enhance credibility of identification strategies. By contrast, Rosenzweig (2000) and Deaton (2010) emphasize that obtaining useful results of randomized controlled economic policy interventions in developing countries requires structural empirical models on behavioral responses to interventions. That kind of economic literature does not, however, relate its arguments to the *ex ante* probability that an effect of interest exists. Moreover, the applications we have in mind are not randomized controlled economic policy interventions in development microeconomics, but those at the macro level where randomized controlled trials are not feasible and IV strategies are potentially useful.

[2]Weakness of instruments refers to the fact that the partial correlation between the instrument(s) and the endogenous explanatory variable(s) approaches zero in absolute value in the population.

[3]Kiviet (2017) also shows that, while a test for relevance of certain regressors in an IV model and a test of overidentifying restrictions may be algebraically equivalent, the maintained hypotheses are different.

2 The Importance of Bayes' Rule – A Quick Reminder

Suppose we are interested in studying the effect of a treatment D (e.g. a region-specific policy intervention) on an outcome Y (e.g. per capita income in the region). Assume that D has been randomly assigned. Consider a regression of the outcome variable Y on the treatment variable D:

$$Y = \gamma_0 + \gamma_1 D + U, \tag{1}$$

where U is an error term with $\mathrm{E}(U) = 0$ and $\mathrm{Cov}(D, U) = 0$. We call eq. (1) the structural equation. Suppose we find that the coefficient on D is significantly different from zero, according to the p-value which is lower than some chosen significance level α (typically, five or one percent). Typically, we interpret such estimate as evidence supporting that D causes Y.[4] What is the probability that in this case D actually affects Y?

To motivate that rigorous theoretical considerations are important for empirical analysis, let us briefly illustrate the importance of the *ex ante* probability that a hypothesis is true or false by recalling Bayes' rule.

Let π denote the *ex ante* probability that the null hypothesis "a causal relationship of D to Y does not exist" ($\gamma_1 = 0$) is true and $1 - \pi$ the *ex ante* probability that it is false (i.e. that the relationship exists). The probability of rejecting the null hypothesis although being true (type I error) is denoted by $\alpha \equiv \Pr\{\gamma_1 = 0 \; reject \,|\, \gamma_1 = 0 \; true\}$. In a regression analysis, the p-value refers to the probability of obtaining a test statistic at least as extreme than the one observed when the null hypothesis is true. The probability of not rejecting the null hypothesis although it is false (type II error), is denoted by $\beta \equiv \Pr\{\gamma_1 = 0 \; not \; reject \,|\, \gamma_1 = 0 \; false\}$.

Using Fig. 1, according to Bayes' rule, the probability that the null hypothesis of no effect is indeed false in our regression example (i.e. the true coefficient on D is non-zero),

[4] This interpretation is, however, misleading. The p-value describes the extent to which the data at hand are compatible with a given null hypothesis and not the probability that the null hypothesis itself is true, see e.g. Berger and Sellke (1987) or Sellke, Bayarri and Berger (2001) for a discussion and numerical examples on this issue.

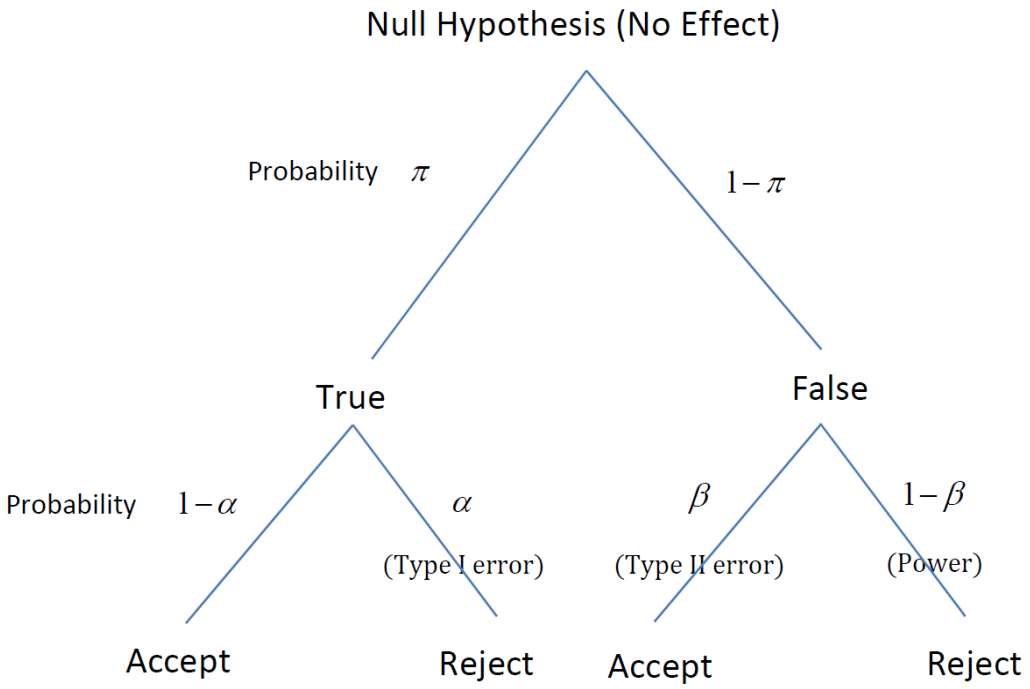

Figure 1: Bayes' rule and hypothesis testing.

given that we reject it, reads as

$$\Pr\{\gamma_1 = 0 \; false \,|\, \gamma_1 = 0 \; reject\} = \frac{(1-\pi)(1-\beta)}{\pi\alpha + (1-\pi)(1-\beta)} = \frac{1}{\frac{\alpha}{R(1-\beta)} + 1} \tag{2}$$

where $R \equiv (1-\pi)/\pi$ can be interpreted as the ratio of false null hypotheses (i.e. the number of instances where there is an effect) to true null hypotheses (i.e. the number of instances where there is no effect) out of the universe of possible hypotheses. It follows from (2) that if, and only if, $R(1-\beta) > \alpha$, then $\Pr\{\gamma_1 = 0 \; false \,|\, \gamma_1 = 0 \; reject\} > 0.5$. In this case, it is more likely that there is indeed an effect given that we found a "significant" coefficient than that the null hypothesis is true (i.e. there is actually no effect despite our seemingly supporting evidence).

As a numerical example, suppose that the *ex ante* probability that D causally affects Y is fifty-fifty ($\pi = 0.5$) which could mean that theoretical arguments in favor of the hypothesis are quite convincing. Also suppose that we require a standard level of significance (i.e. a small probability of a type I error) of five percent and that the probability of

a type II error is only 20 percent. Such a low β requires a sufficiently large sample size, for instance.[5] The power of the test, $1-\beta$, is then 80 percent. With $\alpha = 0.05$, $\beta = 0.2$ and $R = 1$, the probability that the null hypothesis is indeed false, given a coefficient on D that is significantly different from zero, is about 94 percent, according to (2). In this case, empirical support for a causal effect of D on Y, provided that treatment assignment D is properly randomized, is quite strong. If the sample size is small like in many applications in empirical macroeconomics, such that the power of the test drops to 50 percent ($\beta = 0.5$), for $\alpha = 0.05$, we still obtain a probability that the null hypothesis is false given that we reject it based on our regression estimate of almost 91 percent.

However, now suppose the claim that some variable has an effect is based on some weakly substantiated theory, such that the *ex ante* probability π that the null hypothesis is true is 90 percent (i.e. $R = 1/9$). In this case, $\Pr\{\gamma_1 = 0 \; false \,|\, \gamma_1 = 0 \; reject\} = 0.53$. Thus, although a regression coefficient of interest may be significantly different from zero at the five percent level in a seemingly robust way, which many scholars would interpret as strong support that the causal effect of D on Y exists, existence of an effect is not much more probable than its non-existence. If again $\alpha = 0.05$, $\beta = 0.5$, and the *ex ante* probability that D causally affects Y would be a mere five percent ($\pi = 0.95$), then the conclusion on basis of a "significant" coefficient on D that D causally affects Y would only be true with a probability of about one third.

In sum, also apart from endogeneity issues, empirical evidence which lends support to the hypothesis that a causal relationship exists simply based on a "statistically significant" coefficient is quite likely to be misleading if the suggested effect is sufficiently "surprising" to begin with. In fact, the notion of a "surprising" result may just reflect a high *ex ante* probability π (meaning that R is low), giving rise to a likely "false positive" result. We now turn to the case where D is endogenous and argue that even if the *ex ante* probability for $\gamma_1 = 0$ is low, the choice of a "surprising" instrument for D at the first stage can lead to incredible estimates.

[5] In general, the power of the test of the null hypothesis $\gamma_1 = 0$ against the specific alternative that γ_1 takes on a given alternative value depends on the sample size as well as the specific value of γ_1 under the alternative hypothesis.

3 Identification Based on Instrumental Variables

We have just seen that, even in randomized controlled experiments in which the dose of treatment is truly exogenous, empirical estimates may lead to misleading conclusions. In the absence of randomized controlled experiments, the key problem to identify causal effects via regression analysis is potential endogeneity.

More specifically, consider again the regression model of outcome Y on treatment D in eq. (1). Suppose D is not randomly assigned but potentially correlated with the unobserved determinants of Y, i.e. the error term U, so that $\text{Cov}(D, U) \neq 0$. In this case, the OLS estimator of γ_1 is inconsistent. For instance, consider the debate on the factors that fundamentally cause economic exchange or economic growth, like institutions (e.g. the extent of property rights protection and schooling systems) or cultural factors (e.g. religious similarity and common language between trading partners). In a regression analysis, there may be unobserved factors (omitted variables) which affect, for instance, the included regressors (e.g. measures of the quality of institutions) and the dependent variable (e.g. per capita income) at the same time. Applied general equilibrium theory typically suggests that many parameters that capture preferences, production technology (or costs involved in production-related decisions of firms and households) and endowments are candidates for such third factors.

A widely accepted possibility to identify causal effects when $\text{Cov}(D, U) \neq 0$ is to employ an instrumental variables (IV) approach. The IV framework assumes that we have access to an instrument Z that affects the treatment D while it is uncorrelated with the error term of eq. (1), i.e. $\text{Cov}(Z, U) = 0$. The first condition can – to some extent – be verified empirically, while the second condition can only be established based on theoretical arguments.

More specifically, we can express the relationship between the treatment and the instrument as

$$D = \delta_0 + \delta_1 Z + V, \tag{3}$$

where V is an error term with $\mathrm{E}(V) = 0$ and $\mathrm{Cov}(Z, V) = 0$.[6] Eq. (3) is referred to as the first stage or the reduced form for D.

3.1 Instrument Relevance

As argued above, rejecting the null hypothesis that the coefficient δ_1 in eq. 3) is equal to zero does not definitely confirm the first stage relationship between D and Z as the probability that the null is indeed false depends on the *ex ante* probability that Z causally affects D, see eq. (2). A useful check to gain confidence in the relevance of an (excluded) instrument and the existence of the causal relationship posited by the structural equation is to consider the reduced form for the outcome Y. Substituting (3) into (1) we obtain the reduced form for the outcome Y:

$$Y = \gamma_0 + \gamma_1\delta_0 + \gamma_1\delta_1 Z + \gamma_1 V + U = \eta_0 + \eta_1 Z + W, \tag{4}$$

with $\eta_0 \equiv \gamma_0 + \gamma_1\delta_0$, $\eta_1 \equiv \gamma_1\delta_1$, and $W \equiv \gamma_1 V + U$. In fact, since η_1 is the product of the first stage coefficient δ_1 in eq. (3), and the coefficient γ_1 in the structural eq. (1), it can only be non-zero if both γ_1 and δ_1 differ from zero. Angrist and Pischke (2009, p. 213) require that the estimated η_1 is statistically significant, pointing out that "if you can't see the causal relation of interest in the reduced form, it's probably not there".[7]

We will now argue that their statement extends to the consideration of *ex ante* probabilities. Again we can ask the question how likely it is that the coefficient η_1 truly differs from zero if we reject the null hypothesis that it is zero. To see this more clearly, recall $\eta_1 = \delta_1\gamma_1$ to write down the probability that the null hypothesis of no effect of Z on Y is false given that it is rejected as:

$$\Pr(\eta_1 = 0\, false \,|\, \eta_1 = 0\, reject) = \Pr(\delta_1 = 0\, false \wedge \gamma_1 = 0\, false \,|\, \eta_1 = 0\, reject). \tag{5}$$

[6] To streamline the notation the model set up in eq. (1) and (3) omits additional exogenous control variables that are uncorrelated with the error terms. We can think of Y, D and Z as the residuals from regressions of the outcome, the treatment and the instrument on the additional control variables, respectively.

[7] In fact, IV estimates can be severely biased if the instrument is weak, see e.g. Stock et al. (2002) and Kiviet and Niemczyk (2014).

Using the Fréchet inequalities (Fréchet, 1935) we can determine the lower bound of this joint probability as

$$\max[0, \Pr(\delta_1 = 0\, false \,|\, \eta_1 = 0\, reject) + \Pr(\gamma_1 = 0\, false \,|\, \eta_1 = 0\, reject) - 1] \qquad (6)$$

and the upper bound as

$$\min[\Pr(\delta_1 = 0\, false \,|\, \eta_1 = 0\, reject), \Pr(\gamma_1 = 0\, false \,|\, \eta_1 = 0\, reject)]. \qquad (7)$$

To highlight the need for a good theory about both the structural equation *and* the first stage relationship, suppose that $\Pr(\delta_1 = 0\, false \,|\, \eta_1 = 0\, reject) \leq \Pr(\gamma_1 = 0\, false \,|\, \eta_1 = 0\, reject)]$. In this case, we find that $\Pr(\eta_1 = 0\, false \,|\, \eta_1 = 0\, reject) \leq \Pr(\delta_1 = 0\, false \,|\, \eta_1 = 0\, reject)$, according to (7). That is, the *ex ante* probability that the instrument Z affects the outcome Y cannot exceed the *ex ante* probability that the instrument Z affects the treatment variable D, both conditional on rejecting the null hypothesis $\eta_1 = 0$. Thus, even if the *ex ante* probability of a non-zero treatment effect on the outcome is high and we find that η_1 is statistically significant, when the instrument does not affect the treatment, rejection of the null hypothesis $\eta_1 = 0$ does not mean much regarding the effect of the instrumental variable on the outcome. If, in addition, $\Pr(\gamma_1 = 0\, false \,|\, \eta_1 = 0\, reject) \leq 0.5$, which may well be the case if we do not have a convincing theory that supports the treatment effect, then $\Pr(\eta_1 = 0\, false \,|\, \eta_1 = 0\, reject)$ is not even bounded away from zero (!), according to (6).

Thus, when relying on an instrumental variables framework, we should be all the more skeptical about seemingly "creative" choices of instruments that have a low *ex ante* probability that the instrument causes the treatment even if the estimated coefficient is statistically significant in a reduced form equation.

3.2 Instrument Exogeneity

The second condition for a valid instrument is exogeneity, i.e. a candidate instrument must not be correlated with the error term in eq. (1).[8] Practically, this condition rules out the existence of third variables, not included in eq. (1) and (3), that are correlated with both Y and Z.[9] The exogeneity condition has to be substantiated in the context of the specific application at hand using *a priori* arguments. Typically, such arguments are more or less explicitly derived from economic theory. For instance, to estimate demand-side features, determinants of the supply side lend themselves as instruments. Section 4.2 provides a worked example in the context of schooling how rigorous theoretical reasoning can be used to argue both relevance and exogeneity of instrumental variables. Importantly, instrument exogeneity cannot in general be tested empirically.[10]

To illustrate the difficulty to motivate exclusion restrictions, consider the widely accepted practice of using geographic information to construct instruments. Variables capturing regional variation in the proximity to or the availability of specific facilities / institutions (e.g. schools, hospitals, retail stores) are often used as instruments when the goal is to asses the causal effect of these facilities / institutions on economic outcomes such as earnings, employment or per capita income, see e.g. Card (1995), Neumark, Zhang and Ciccarella (2008) or Becker and Woessmann (2009) for prominent examples.[11] Using geographic variables as instruments is not innocuous, though, because they may violate the exclusion restriction if other locational (economic or cultural) factors are not controlled for. In fact, the location of settlements, agglomerations or industries is typically not random but a result of political, geographic or climatic factors that may well have influenced both instrument and outcome or may have affected third variables

[8] It is well known that violation of the exclusion restriction could imply that an IV coefficient is more biased than its OLS counterpart, see e.g. Hahn and Hausman (2005) and Kiviet and Niemczyk (2014).

[9] More precisely, this requirement has to hold after partialling out all other observed covariates, see footnote 6.

[10] When multiple instruments are available one could in principle conduct an overidentification test to test whether the instruments are valid. However, they still require to maintain an *a priori* exogeneity condition for the just-identified model, which often seems implausible in applied work, see e.g. Murray (2006), Angrist and Pischke (2009, ch. 4) and Kiviet (2017) for a discussion.

[11] Also other disciplines use geographic information as instruments, see e.g. the survey by Garabedian et al. (2014) on the medical literature.

correlated with instrument and outcome. One may wish to determine the probability that any form of spatial interdependence relating to the instrument is irrelevant for the outcome using Bayes' rule. Unfortunately, we cannot illustrate this probability by invoking Bayes' rule, as this would require some hypothesis test that is not available for the exclusion restriction. Therefore, theory to justify the exclusion restriction is salient. Moreover, careful applications of IV techniques usually provide additional empirical evidence to support the hypothesis of instrument exogeneity, which, however, cannot be interpreted as a formal test.

We briefly discuss some prominent, controversially received literature that employs geographic distance as instrument. We start with the study of Becker and Woessmann (2009) who suggest that Protestant regions in 19th century Prussia had higher literacy rates and therefore higher per capita income compared to Catholic regions. The instrumental variable for Protestantism is distance to Wittenberg, the home town of Martin Luther. Their basic argument is that "distance to Wittenberg is indeed unrelated to a series of proxies for economic and educational development before 1517, including the pre-Luther placement of schools, universities, monasteries, and free imperial and Hanseatic cities and urbanization" (Becker and Woessmann, 2009, p. 532). However, Edwards (2017) criticizes that they do not take into account systematic regional heterogeneity. He shows that regional effects are empirically important and that the original results of Becker and Woessmann (2009) do not persist after taking into account regional heterogeneity.[12]

Neumark *et al.* (2008) study the effect of Wal-Mart store openings on employment and earnings in the retail sector in U.S. counties. They use distance to the Wal-Mart headquarters interacted with time to instrument store openings. The IV strategy exploits that Wal-Mart expanded from a local chain store to a national one by spreading

[12] Boppart et al. (2013) and Boppart, Falkinger and Grossmann (2014) employ a similar idea to examine the role of variation in the population share of Protestants at the district level for educational test results of military conscripts in 19th century Switzerland. They use the distance to the historical centers of Protestantism (shorter distance of a district to Zurich and Geneva) as instrument for Protestantism. To mitigate the concern that the instrument violates the exclusion restriction, they also control for geographical and economic factors like altitude, population density, and the (closest) proximity to one of the six major Swiss cities.

out geographically to counties farther away. Thus, in a given year, the distance of a county from Wal-Mart's headquarters predicts the probability that a new store is opened. Basker (2007) questions the validity of the instrument and provides empirical evidence suggesting that the instrument may be correlated with third factors that also affect the performance of local labor markets. More specifically, Wal-Mart's headquarters lie in a rather rural area in the central south of the U.S. while the metropolitan areas are located closer to the coasts. Thus, counties at the coasts differ systematically from counties in the center not just in the timing of the opening of Wal-Mart stores but also in their population density, industry structure and other economically relevant characteristics.

4 Rigorously Founding Identification Strategies

4.1 Challenges

Summarizing the above, finding a good excluded instrument is notoriously difficult for at least two reasons. First, for its relevance, an excluded instrument should have a high *ex ante* probability of being partially correlated with the endogenous regressor in an important way, as argued above by exploiting Bayes' rule and the Fréchet inequalities. Otherwise, the instrument relevance assumption may be violated, leaving the IV estimate inconsistent. Second, the exclusion restriction is not easily justified. As known by every applied general equilibrium theorist, the exogenous factors of a theoretical model (i.e. the parameters characterizing technology, preferences, and endowments) typically pop up in many endogenous variables, except under special assumptions. In a theoretical model, special assumptions may be justified for the purpose of highlighting a specific economic mechanism. However, for instrument exogeneity, it is necessary to argue quite generally that an exogenous factor is not correlated with an endogenous variable that *cannot* be controlled for in the empirical model. Doing so typically requires rigorous theoretical analysis that advises us which explanatory variables of interest shall be treated as endogenous and which are candidate instruments.

The long-standing debate on formalization of economic theory has led to the con-

clusion that rigorous theoretical foundations are required to show that an intuition is consistent with mathematically proven conclusions derived under explicit, transparent assumptions. It has even provided economists with confidence that testable, theoretical hypotheses are typically better founded in economic research than in other social sciences. However, when it comes to justify that an instrument is partially correlated with an endogenous regressor and at the same time fulfills the exclusion restriction, seemingly intuitive (verbal) reasoning is the standard in many successfully published empirical applications. We will now illustrate how economic theory can rather be used to *rigorously* derive an IV strategy in the context of schooling and its effect on per capita income.

4.2 Illustration: The Macroeconomic Effects of Schooling

Estimating the causal effect of the amount of human capital in an economy (typically measured by the average years of schooling or a measure of the population average of cognitive skills) on aggregate income and/or investment in physical capital is a long-standing issue.[13] Finding an excluded instrument requires a dynamic general equilibrium model that suggests an observable exogenous factor affecting investment and income only through its effect on human capital and not independently of it. We now demonstrate how such a candidate instrument can be derived.

4.2.1 Theoretical Set Up

To fix ideas and illustrate the challenge IV estimations based on aggregate data may impose, consider the following perfectly competitive environment in continuous time. Suppose (per capita) income y is equal to output of a unit mass of identical firms producing a single consumption good, chosen as numeraire. Production function f depends on the (per capita) stocks of physical and human capital devoted to production, denoted by k^Y and h^Y, respectively. We specify

$$y = f(k^Y, h^Y) = A \cdot \left(k^Y\right)^{\alpha} \cdot (h^Y)^{1-\alpha}, \tag{8}$$

[13] See e.g. Barro (1991), Bils and Klenow (2000) and Cohen and Soto (2007), among others.

$A > 0$, $0 < \alpha < 1$. Physical capital depreciates at constant rate $\delta_K > 0$. The (per capita) stock of human capital in the economy, h, accumulates according to[14]

$$\dot{h} = B \cdot [(1 - s) \cdot h]^{\eta} - \delta_H \cdot h, \tag{9}$$

$B > 0$, $0 < \eta \leq 1$, where $\delta_H > 0$ is the human capital depreciation rate and $1 - s$ is the (average) fraction of time devoted to education. The case of constant returns in educational production, $\eta = 1$, is treated in the seminal paper on endogenous growth theory by Lucas (1988). It implies that human capital grows without bound, a feature that has been criticized (e.g. Temple, 2001). We will therefore focus on the case of decreasing returns in educational production, $\eta < 1$, and relegate the discussion of the case $\eta = 1$ to the appendix.[15] We follow Lucas (1988) to allow for human capital externalities. That is, total factor productivity (TFP), A, may depend on the (per capita) human capital level h. However, because a single firm has mass zero, the relationship is not taken into account by firms. Thus, they take A as given when choosing inputs to maximize profits. We specify

$$A = a \cdot h^{\beta}, \tag{10}$$

$a > 0$, $\beta \geq 0$.[16] Parameter a may be viewed as capturing historically rooted factors affecting TFP.

There is an infinitely living, representative household with unit time endowment. It lends its non-human assets, k, to the representative firm and inelastically supplies human capital that is not devoted to education (i.e. fraction s of amount h) to the labor market. Thus, $k^Y = k$ and $h^Y = sh$. Using these equilibrium conditions and (10) in (8),

[14] We denote by $\dot{x}(t) \equiv \mathrm{d}x(t)/\mathrm{d}t$ the derivative of a variable x with respect to time t. The time index is omitted whenever this does not lead to confusion.

[15] In the appendix, we argue that the derived identification strategy for determining the causal relation from human capital to per capita income for the case where $\eta = 1$ is similar to the case where $\eta < 1$; see Proposition A.1 and its discussion. We also argue, however, that identification of the causal effect of schooling on physical capital investment may be impossible when based on assumption $\eta = 1$.

[16] The reader may excuse that we use α and β as production elasticities in this section while they denoted the probability of a type I and type II error in section 2, respectively.

per capita income can be written as

$$y = ak^{\alpha}s^{1-\alpha}h^{1-\alpha+\beta}. \tag{11}$$

The amount of non-human assets accumulates according to

$$\dot{k} = wsh + rk - c, \tag{12}$$

where w denotes the wage rate per unit of human capital and r denotes the interest rate net of depreciation.

Let $c(t)$ denote the household's consumption of the numeraire good at time t. The household chooses the consumption path and time allocation variable, s, to maximize intertemporal utility

$$\int_0^{\infty} u(c(t))e^{-\rho t}\mathrm{d}t, \text{ with } u(c) = \begin{cases} \frac{c^{1-\sigma}-1}{1-\sigma} \text{ for } \sigma \neq 1 \\ \log c \text{ otherwise} \end{cases} \tag{13}$$

subject to (9) and (12), $\rho > 0$, $\sigma > 0$. Initial values of stock variables, $h(0) = h_0 > 0$ and $k(0) = k_0 > 0$, are given.

4.2.2 Equilibrium Analysis

The definition of an equilibrium is relegated to the appendix. We focus the equilibrium analysis on the long run and look for a balanced growth equilibrium (BGE), where all variables grow at a constant (possibly zero) rate. Long run values are denoted by superscript (*).

Proposition 1. *Suppose $\eta < 1$. There exists a BGE, such that the key variables are stationary and given by*

$$s = \frac{\rho + \delta_H(1-\eta)}{\eta\delta_H} \equiv s^*, \tag{14}$$

$$h = \left(\frac{B(\rho+\delta_H)^{\eta}}{\eta^{\eta}(\delta_H)^{1+\eta}}\right)^{\frac{1}{1-\eta}} \equiv h^*, \tag{15}$$

$$k = \left(\frac{\alpha a}{\rho + \delta_K}\right)^{\frac{1}{1-\alpha}} (h^*)^{\frac{1-\alpha+\beta}{1-\alpha}} s^* \equiv k^*, \tag{16}$$

$$y = a^{\frac{1}{1-\alpha}} \left(\frac{\alpha}{\rho + \delta_K}\right)^{\frac{\alpha}{1-\alpha}} (s^*)^{\alpha} (h^*)^{\frac{1-\alpha+\beta}{1-\alpha}} \equiv y^*. \tag{17}$$

Proof. See appendix. ■

Since the long run level of human capital (h^*) is stationary if $\eta < 1$ (i.e. under decreasing returns in educational production), there is no TFP growth in BGE. Thus, also the level of physical capital is stationary and the long run investment in physical capital, denoted by I^*, is equal to the amount of physical capital that is depreciating, $I^* \equiv \delta_K k^*$. Now consider the impact of an increase in the marginal benefit B to devote time to education. On the one hand, an increase in B provides an incentive to shift the time allocation towards education (as $\frac{\partial^2 \dot{h}}{\partial(1-s)\partial B} > 0$, according to (9)). On the other hand, however, the resulting increase in the (long run) level of human capital raises to the same extent the marginal (opportunity) costs of doing so in the form of foregone labor income. In sum, in BGE, the long run value of the time allocation variable, s^*, is independent of B. Similarly, an increase in TFP parameter a, by raising the long run wage rate, raises both the benefit and the opportunity costs of schooling alike. None of the long run values in (14)-(17) are affected by initial levels of stock variables, h_0 and k_0. Consequently, the following result holds.

Corollary 1. *Suppose $\eta < 1$. An increase in educational productivity, B, affects long run investment in physical capital, I^*, and long run per capita income, y^*, only through a change in the long run level of human capital, h^*, but not through a behavioral response on the long run level of the time allocation variable, s^*. An increase in TFP parameter a directly affects per capita income (y^*) and investment (I^*) without affecting s^* or h^*. Initial levels of human and physical capital, h_0 and k_0, respectively, do not directly affect s^*, h^*, I^*, y^*.*

Most importantly, Corollary 1 suggests that the productivity in both human capital formation and final output production do not affect the (possibly unobserved) allocation

of human capital between labor supply and educational production. As shown in Grossmann *et al.* (2015), the results are robust to changing the human capital formation process to reflect privately financed (and, possibly, publicly subsidized) wage costs of teachers rather than time opportunity costs.

4.2.3 Structural Derivation of IV Strategies

Based on Corollary 1, we can now discuss potential IV strategies. Adding subscript i to denote regions (or countries), we use (17) to write

$$\log y_i^* = \gamma_0 + \gamma_1 \log h_i^* + \gamma_2 \log a_i + U_i, \tag{18}$$

where $\gamma_0 \equiv \frac{\alpha}{1-\alpha} \log \alpha$, $\gamma_1 \equiv \frac{1-\alpha+\beta}{1-\alpha}$, $\gamma_2 \equiv \frac{1}{1-\alpha}$, and $U_i \equiv \alpha \log s_i^* - \frac{\alpha}{1-\alpha} \log (\rho_i + \delta_K)$. Based on (18), we may estimate the causal effect of an increase in the level of human capital on (the log of) per capita income in a sample of regions or countries.[17] We assume that elasticities α, β, η, and depreciation rates δ_K, δ_H, are the same for all regions, whereas productivity parameters a_i, B_i, and the time preference rate ρ_i are interpreted as being institutionally or culturally rooted and thus may differ across regions. ρ_i may be viewed as capturing patience. Doepke and Zilibotti (2008, 2013) strongly argue that patience is indeed a fundamental determinant of economic development that can explain the regional variation in per capita income. According to (14) and (17), the time preference rate ρ_i affects the error term directly and via the long run value of the time allocation variable, s_i^*, which is unobservable. It also affects h_i^*, according to (15). Consequently, if ρ_i were unobservable and differed across regions or countries, h_i^* would be correlated with the error term and thus an OLS estimate of γ_1 would be biased.

By contrast, according to Corollary 1, the productivity in human capital formation, B_i, critically affects h_i^*, without affecting s_i^*, i.e. is unrelated to the error term, U_i. The theoretical model thus suggests that an appropriate measure of B_i may serve as a valid excluded instrument to address the mentioned endogeneity problem. The analogous

[17] It is noteworthy that we must not control for the capital stock, since k^* is proportionally related to h^*, according to (16). Otherwise, the rank condition would be violated.

reasoning applies to considering long run physical capital investment, $I_i^* = \delta_K k_i^*$, rather than per capita income as dependent variable.

Application 1: Hanushek and Woessmann (2012) Hanushek and Woessmann (2012) are interested in the effect of schooling on the growth rate of per capita income between 1960 and 2000. They estimate the following cross-country regression:

$$\log y_i^* - \log y_{i0} = \gamma_0 + \gamma_1 H_i + \gamma_2 \log y_{i0} + \mathbf{x}_i \boldsymbol{\gamma}_x + U_i, \tag{19}$$

where H_i is a measure of contemporaneous cognitive math and science skills (the average test score across pupils in primary to lower secondary education in country i, based on international student achievement tests and averaged over an extended period), y_i^* and y_{i0} denote i's GDP per capita in the year 2000 and 1960, respectively, $\mathbf{x}_i$ are other controls (with coefficients $\boldsymbol{\gamma}_x$), and U_i is the error term.

Hanushek and Woessmann (2012) employ the presence of an external exit exam (a country-specific dummy variable), Z_i, hypothesized to positively affecting test scores H_i, as excluded instrument in a 2SLS regression analysis. They argue that: "External exit exam systems are a device to increase accountability in the school system that has been repeatedly shown to be related to better student achievement" (p. 283). Viewing H_i as a measure of $\log h_i^*$ and excluded instrument Z_i as a measure of educational productivity B_i in the proposed theoretical model, the IV strategy correctly identifies γ_1, according to Corollary 1, (18) and (19). Moreover, we may view initial GDP level y_{i0} in estimated equation (19) as being related to the historically rooted TFP parameter a_i in (18) that entered the theoretical model. In this case, because of the term $-\log y_{i0}$ on the left-hand side of (19), the sign of γ_2 could be positive or negative. The estimated γ_2 in Hanushek and Woessmann (2012) is negative (typically interpreted as neoclassical convergence force). Most importantly and reassuringly in light of our theoretical considerations, the estimated γ_1 is positive and highly significant.

In sum, even though their empirical model is not structurally derived explicitly, both the identification strategy and estimates of Hanushek and Woessmann (2012) are *as if*

it is, giving much credibility to their results.

Application 2: Tabellini (2010) Also based on a 2SLS regression analysis, Tabellini (2010) argues that, in addition to variation in human capital levels, differences in culturally transmitted values like "generalized trust" that inhabitants within a region generally have on average towards other people can explain the variation of per capita income across European regions. The hypothesis is theoretically well-founded (Putnam, 1993; Zak and Knack, 2001) and gained empirical support (e.g. Knack and Keefer, 1997), albeit causal empirical inference has been missing. To fill this gap, Tabellini (2010) estimates the structural equation

$$\log y_i^* = \gamma_0 + \gamma_1 H_i + \gamma_2 Culture_i + \mathbf{x}_i \boldsymbol{\gamma}_x + U_i, \tag{20}$$

where y_i^* is the level of per capita income averaged over the period 1995-2000, H_i is the gross enrolment rate in primary and secondary schools in 1960, $Culture_i$ denotes the contemporaneous cultural variable (in 1999-2000), $\mathbf{x}_i$ denotes other controls, and U_i is the error term. Tabellini (2010) uses the literacy rate at the end of the 19th century and political institutions in the mid-19th century and earlier as instrumental variables, Z_i, to address the potential endogeneity of culture.

We will now employ our theoretical considerations to uncover the assumptions under which Tabellini (2010) correctly identifies the effects of cultural variation across regions on per capita income. First, viewing H_i as a measure of $\log h_i^*$ on the right-hand side of (18) without instrumenting human capital requires that H_i is not correlated with the error term. That would hold if we viewed H_i as measure for the productivity in human capital formation in the theoretical model, B_i. Measures of $Culture_i$ like generalized trust are likely to be endogenous, because they may be related to patience as conceptualized by the time preference rate ρ_i, which is contained in the error term $U_i = \alpha \log s_i^* - \frac{\alpha}{1-\alpha} \log(\rho_i + \delta_K)$ in (18). (Recall that also s_i^* depends on ρ_i.) Generalized trust or other cultural measures may also be correlated with productivity parameters a_i and B_i that, in the theoretical model, are unrelated to U_i. Thus, allowing $Culture_i$ to

be related to a_i, B_i, $\mathbf{x}_i$, ρ_i, excluded instruments may be related to a_i, B_i and $\mathbf{x}_i$, but must not be related to ρ_i.

In particular, suppose that TFP parameter a_i (capturing historical roots) is determined by region i's literacy rate and political institutions in the distant past, denoted by h_{0i} and P_i, respectively. Also assume that h_{0i} and P_i are uncorrelated with patience, ρ_i, that is conceptionally separated from TFP parameter, a_i, in our model. Also recall that we assumed B_i to be captured by past enrolment rates, H_i. Tabellini (2010) linearly regresses $Culture_i$ on h_{0i}, P_i, H_i and $\mathbf{x}_i$. Since a_i, B_i, and the initial human capital level, h_{0i}, are not contained in the theoretically derived error term U_i, instruments $Z_i = (h_{0i}, P_i)$ fulfill the exclusion restrictions on grounds of our theoretical model under the assumption.

Tabellini (2010) also estimated a reduced form regression by leaving out $Culture_i$ in (20) and using $Z_i = (h_{0i}, P_i)$ instead as controls, in addition to H_i and $\mathbf{x}_i$. He indeed finds significant coefficients on historical variables Z_i in such an estimation, supporting the instrument relevance assumption. Reassuringly, the estimated coefficients of the instrumented variable $Culture_i$ in the structural estimation (20) suggest rejecting the null hypothesis that $\gamma_2 = 0$.

The application illustrates the many steps and explicit assumptions necessary to rigorously motivate an identification strategy and attribute high *ex ante* probabilities to the non-zero effects. As a further caveat, our theoretical considerations that justify the identification strategies of both Hanushek and Woessmann (2012) and Tabellini (2010) presume that the economy is in steady state, at least approximately. At least, such assumption becomes visible when rigorously founding IV approaches.[18]

[18] Having identified the critical assumptions, we could numerically examine transitional dynamics of the theoretical model to check whether the exclusion restrictions are largely supported also off steady state.

5 Concluding Remarks

In this paper, we demonstrated how a structural approach could help to obtain credible results of IV estimations. First, a rigorous theoretical foundation advises the researcher on candidates for valid, excluded instruments. Second, it enhances the *ex ante* probability that a hypothesized effect exists both at the first stage (instrument relevance) and at the estimation of the structural equation. According to Bayes' rule, the *ex ante* probability critically determines the conditional probability that the null hypothesis (of no effect) is indeed false given that it is rejected at conventionally employed significance levels. Conventional robustness checks are not sufficient to prevent researchers from jumping to erroneous conclusions, if the theoretical reasoning of a hypothesized causal effect is weak to begin with.

In the empirical macro-development literature, new data cannot be readily generated and are often of aggregate nature. Thus, replication analysis typically is more challenging than in medical and pharmaceutical research where the same experiment can just be redone. Replication studies are also (too) rarely published in highly reputable outlets even when results contradict the original ones. "False positive" results could thus become conventional wisdom for an extensive period.[19]

As a result, researchers may have distorted incentives that could foster data mining (i.e. pre-testing) in the search of seemingly relevant, excluded instrumental variables. Its choices may be *ex post* rationalized by some intuitive reasoning as a substitute for elaborate theoretical considerations. It is well-known, however, again based on Bayesian arguments, that pre-testing variables and fishing out the significant ones in regression analysis leads to invalid inference (e.g. Leamer, 1978). Improved standards for empirical analyses that include structural modeling could also mitigate the well-known "publication bias" that incentivizes to generate significant coefficient estimates by testing *ex ante*

[19] In the Online-Appendix, we discuss two widely-received IV studies that exploit aggregate data and have been falsified by later work. See Albouy (2012) on Acemoglu *et al.* (2001), who study the effect of higher institutional quality (property rights protection) on economic development, and Spring and Grossmann (2015) on Guiso *et al.* (2009), who examine the effect of closer cultural proximity on international goods trade. We discuss the approaches from a theoretical point of view and suggest what can be learnt from the replication studies in future empirical research.

improbable hypotheses.

Appendix

This Appendix provides the equilibrium analysis of the model developed in section 4.2. We first define the equilibrium and prove Proposition 1 (applying to $\eta < 1$) and then discuss IV strategies for the case of endogenous long run growth ($\eta = 1$).

Definition 1. *A market equilibrium consists of time paths for the quantities* $\{k_t, k_t^Y, h_t^Y, h_t, c_t, y_t\}_{t=0}^{\infty}$, *time allocation variable* $\{s_t\}_{t=0}^{\infty}$, *and prices* $\{w_t, r_t\}_{t=0}^{\infty}$ *such that*

(i) the representative household maximizes intertemporal welfare (13) subject to (9), (12) and non-negativity constraints;

(ii) firms produce output according to (8) and maximize profits, $y-(r+\delta_K)k^Y-wh^Y$, *by taking total factor productivity* $A = ah^{\beta}$ *as given;*

(iii) factor markets clear, i.e. $h_t^Y = s_t h_t$ *and* $k_t^Y = k_t$.[20]

Proof of Proposition 1. The current-value Hamiltonian corresponding to the optimization problem of the household (condition (i) in Definition 1) is given by

$$\mathcal{H} = \frac{c^{1-\sigma}-1}{1-\sigma} + \mu \cdot [B(1-s)^{\eta}h^{\eta} - \delta_H h] + \lambda \cdot (rk + wsh - c),$$

where μ and λ are multipliers (co-state variables) associated with (9) and (12), respectively. Necessary optimality conditions are $\partial\mathcal{H}/\partial c = \partial\mathcal{H}/\partial s = 0$ (control variables), $\dot{\mu} = \rho\mu - \partial\mathcal{H}/\partial h$, $\dot{\lambda} = \rho\lambda - \partial\mathcal{H}/\partial k$ (state variables), and the corresponding transversality conditions,

$$\lim_{t\to\infty} \mu_t e^{-\rho t} h_t = \lim_{t\to\infty} \lambda_t e^{-\rho t} k_t = 0. \tag{21}$$

Thus,

$$\lambda = c^{-\sigma}, \tag{22}$$

[20] According to condition (ii), final output reads as $y = (r+\delta_K)k^Y + wh^Y$. According to (12) and condition (iii), $\dot{k}^Y = \dot{k} = rk + wh^Y - c$. Thus, $\dot{k} = y - c - \delta_K k$ or $y = c + I$, with $I \equiv \dot{k} + \delta_K k$ being equal to gross investment. Hence, also the final goods market clears (Walras' law).

$$\mu\eta B(1-s)^{\eta-1}h^{\eta} = \lambda w h, \tag{23}$$

$$\frac{\dot{\mu}}{\mu} = \rho - \eta B\left(1-s\right)^{\eta} h^{\eta-1} + \delta_H - \frac{\lambda}{\mu} w s, \tag{24}$$

$$\frac{\dot{\lambda}}{\lambda} = \rho - r. \tag{25}$$

Differentiating (22) with respect to time and using (25), we obtain Euler equation

$$\frac{\dot{c}}{c} = \frac{r-\rho}{\sigma}. \tag{26}$$

Combining (23) and (24), we find

$$\frac{\dot{\mu}}{\mu} = \rho - B\eta\left(1-s\right)^{\eta} h^{\eta-1} + \delta_H - B\eta(1-s)^{\eta-1}h^{\eta-1}s. \tag{27}$$

Under perfect competition, the wage rate, w, equals the marginal product of human capital and the user costs of capital, $r+\delta_K$, equals the marginal product of physical capital (thus, in equilibrium, the profit of the representative firm is zero), i.e.

$$w = (1-\alpha)A\left(\frac{k^Y}{h^Y}\right)^{\alpha}, \tag{28}$$

$$r = \alpha A\left(\frac{h^Y}{k^Y}\right)^{1-\alpha} - \delta_K. \tag{29}$$

We search for a BGE and suppose $\dot{c} = \dot{h} = \dot{\mu} = 0$. Setting $\dot{c} = 0$ in (26) implies $r = \rho$. Setting $\dot{h} = 0$ in (9) and solving for h we find

$$B(1-s)^{\eta}h^{\eta-1} = \delta_H. \tag{30}$$

Substituting (30) into (27) and setting $\dot{\mu} = 0$ implies (14). Substituting (14) into (30) and solving for h confirms (15). Using $r = \rho$ and $A = ah^{\beta}$ in (29) implies (16). Also use (15) in (11) to confirm (17).

Finally, we have to show that a BGE with equilibrium values (14)-(17) exists. First, note that y^* in (17) is stationary. Second, substitute $k^Y = k^*$, $h^Y = sh^*$ and $A = a(h^*)^{\beta}$

into (28) to confirm that the long run wage rate is stationary. The long run consumption level can be obtained residually from setting $\dot{k} = 0$ in (12) and using both $k = k^*$ and $h = h^*$. It is stationary as well. Also note that, since μ, h, λ and k are stationary in the long run, the transversality conditions (21) hold. ■

Proposition A.1. *Denote the growth rate of a variable x by $g_x \equiv \dot{x}/x$. Suppose that $B > \rho$ and $\eta = \sigma = 1$.*[21] *There exists an interior BGE, such that*

$$g_h = B - \rho - \delta_H \equiv \hat{g}_h, \tag{31}$$

$$s = \frac{\rho}{B} \equiv \hat{s}, \tag{32}$$

$$g_k = g_y = g_c = \frac{1 - \alpha + \beta}{1 - \alpha} \hat{g}_h \equiv \hat{g}_y, \tag{33}$$

and long run levels of $\tilde{h} \equiv he^{-\hat{g}_y}$, $\tilde{k} \equiv ke^{-\hat{g}_y}$, $\tilde{y} \equiv ye^{-\hat{g}_y}$ are stationary but indeterminate (i.e. generally depend on initial conditions, h_0 and k_0).

Proof. Using $\eta = 1$ in (23) implies that $\lambda/\mu = B/w$. Thus, $g_w = g_\mu - g_\lambda$ and, according to (24), $g_\mu = \rho + \delta_H - B$. Using $\sigma = 1$ in (26), we can write $r = g_c + \rho$. Thus, according to (25), $-g_\lambda = g_c$. Hence,

$$g_w = g_\mu - g_\lambda = \rho + \delta_H - B + g_c. \tag{34}$$

We seek for a BGE where r and s are stationary and both labor and non-labor income grow at the same rate as consumption, which is a candidate for a steady state, according to (12). If s is stationary, the growth rate of labor income reads as $g_w + g_h$, which we set equal to g_c in search for a BGE. Combining $g_h = g_c - g_w$ with (34) confirms (31). Using $\eta = 1$ in (9), we have $g_h = B(1 - s) - \delta_H$. Combining the latter with (31) confirms (32). Note that $\hat{s} < 1$ if $B > \rho$. Finally, according to (11) and $\dot{s} = 0$, we have $g_y = \alpha g_k + (1 - \alpha + \beta) g_h$. Substituting both $g_h = \hat{g}_h$ and $g_y = g_k$ into the latter equation

[21] Lucas (1988) focussed on $\eta = 1$ and showed that endogenous long run growth may emerge. For simplicity, we focus on the standard case where the coefficient of relative risk aversion $\sigma = 1$. Applying L'Hôpital's rule, we have $\lim_{\sigma \to 1} \frac{c^{1-\sigma} - 1}{1 - \sigma} = \log c$. Assuming a logarithmic instantaneous utility function does not affect our main conclusions.

confirms (33). It is also easy to check that the transversality conditions (21) hold. The reminder of the proof is obvious. ■

Possible IV Strategy Based on Proposition A.1. If the case $\eta = 1$ rather than $\eta < 1$ were the correct model specification, we should regress the growth rate of per capita income on the growth rate of human capital in a panel analysis with countries or regions as observational units. According to (31) and (33), we could again use a measure of educational productivity B as excluded instrument in a 2SLS approach, since in BGE a change in B affects per capita income growth rate $\hat{g}_y$ only through its impact on $\hat{g}_h$.

However, when it comes to the effect of schooling on investment in physical capital, the appropriate IV strategy is less clear when considering the case where $\eta = 1$. Gross investment reads as $I \equiv \dot{k} + \delta_K k$, i.e. $I = (g_k + \delta_K)k$. Since $g_k = \hat{g}_y$ and $k = \tilde{k}e^{\hat{g}_y}$ in BGE, long run gross investment can be written as $\hat{I} \equiv (\hat{g}_y + \delta_K)\tilde{k}e^{\hat{g}_y}$. Thus, we have

$$\log \hat{I} = \log(\hat{g}_y + \delta_K) + \log \tilde{k} + \hat{g}_y. \tag{35}$$

First, there is the difficulty that the effect of a change in long run growth rate $\hat{g}$ on both $\hat{I}$ and $\log \hat{I}$ is non-linear. A first-order approximation may be acceptable to deal with that problem, such that (35) may be written as

$$\log \hat{I} \simeq \gamma_0 + \gamma_1 \hat{g}_y + U, \tag{36}$$

where U is the error term that includes $\log \tilde{k}$, γ_0 is a constant, and γ_1 is the coefficient of interest. More fundamentally, however, an increase in B does not only affect physical capital investment via

$$\hat{g}_y = \frac{(1 - \alpha + \beta)(B - \rho - \delta_H)}{1 - \alpha} \tag{37}$$

(use (31) and (33)), but also via the detrended (and stationary) physical capital stock in BGE, $\tilde{k}$. Identification is thus non-obvious and potentially impossible.

References

[1] Acemoglu, Daron, Simon H. Johnson and James A. Robinson (2001). The Colonial Origins of Comparative Development: An Empirical Investigation, *American Economic Review* 91, 1369-1401.

[2] Albouy, David Y. (2012). The Colonial Origins of Comparative Development: An Investigation of the Settler Mortality Data, *American Economic Review* 102, 3059-3076.

[3] Angrist, Joshua D. and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton.

[4] Barro, Robert J. (1991). Economic Growth in a Cross Section of Countries, *Quarterly Journal of Economics* 106, 407-443.

[5] Barro, Robert J. and Rachel M. McCleary (2003). Religion and Economic Growth, *American Sociological Review* 68, 760-781.

[6] Barro, Robert J. and Rachel M. McCleary (2006). Religion and Political Economy in an International Panel, *Journal for the Scientific Study of Religion* 45, 149-175.

[7] Basker, Emek (2007). When Good Instruments Go Bad: A Reply to Neumark, Zhang, and Ciccarella, Working Paper No. 0706, Department of Economics, University of Missouri.

[8] Becker, Sascha O. and Ludger Woessmann (2009). Was Weber Wrong? A Human Capital Theory of Protestant Economic History, *Quarterly Journal of Economics* 124, 531-596.

[9] Berger, James O. and Thomas Sellke (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association* 82, 112-122.

[10] Bils, Mark and Peter J. Klenow (2000). Does Schooling Cause Growth?, *American Economic Review* 90, 1160-1183.

[11] Boppart, Timo, Josef Falkinger, Volker Grossmann, Ulrich Woitek and Gabriela Wüthrich (2013). Under Which Conditions Does Religion Affect Educational Outcomes?, *Explorations in Economic History* 50, 242–266.

[12] Boppart, Timo, Josef Falkinger, Volker Grossmann (2014). Protestantism and Education: Reading (the Bible) and Other Skills, *Economic Inquiry* 52, 874–895.

[13] Card, David (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling, in L. Christofides, E.K. Grant and R. Swindinsky (eds.), *Aspects of Labour Economics: Essays in Honour of John Vanderkamp*, Toronto: University of Toronto Press.

[14] Cohen, Daniel and Marcelo Soto (2007). Growth and Human Capital: Good Data, Good Results, *Journal of Economic Growth* 12, 51-76.

[15] Deaton, Agnus (2010). Instruments, Randomization, and Learning about Development, *Journal of Economic Literature* 48, 424-455.

[16] Doepke, Matthias and Fabrizio Zilibotti (2008). Occupational Choice and the Spirit of Capitalism, *Quarterly Journal of Economics* 123, 747–793.

[17] Doepke, Matthias and Fabrizio Zilibotti (2013). Culture, Entrepreneurship, and Growth, in: Philippe Aghion and Steven N. Durlauf (Eds.), *Handbook of Economic Growth*, Vol. 2, 1-48.

[18] Edwards, Jeremy (2017). Did Protestantism Promote Economic Prosperity via Higher Human Capital?, CESifo Working Paper No. 6646, CESifo, Munich.

[19] Fehr, Ernst and Jean-Robert Tyran (2001). Does Money Illusion Matter?, *American Economic Review* 91, 1239-1262.

[20] Fehr, Ernst and Jean-Robert Tyran (2008). Limited Rationality and Strategic Interaction: The Impact of the Strategic Environment on Nominal Inertia, *Econometrica* 76, 353-394.

[21] Fréchet, Maurice (1935). Généralisations du théorème des probabilités totales, *Fundamental Mathematica* 25, 379–387.

[22] Garabedian, Laura F., Paula Chu, Sengwee Toh, Alan M. Zaslavsky and Stephen B. Soumerai (2014). Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research, *Annals of Internal Medicine* 161, 131-138.

[23] Grossmann, Volker, Thomas M. Steger and Timo Trimborn (2015). Quantifying Optimal Growth Policy, *Journal of Public Economic Theory* 18, 451-485.

[24] Guiso, Luigi, Paola Sapienza and Luigi Zingales (2006). Does Culture Affect Economic Outcomes, *Journal of Economic Perspectives* 20, 23-48.

[25] Guiso, Luigi, Paola Sapienza and Luigi Zingales (2009). Cultural Biases in Economic Exchange?, *Quarterly Journal of Economics* 124, 1095-1131.

[26] Hahn, Jinyong and Jerry Hausman (2005). Estimation with Valid and Invalid Instruments, *Annales d'Économie et de Statistique* 79/80, 25-57.

[27] Hanushek, Eric A. and Ludger Woessmann (2012). Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation, *Journal of Economic Growth* 17, 267-321.

[28] Ioannidis, John P.A. (2005). Why Most Published Research Findings Are False, *PLoS Medicine* 2 (8), 696-701.

[29] Kiviet, Jan F. and Jerzy Niemczyk (2014). On the Limiting and Empirical Distribution of IV Estimators When Some of the Instruments are Actually Endogenous, in: Y. Chang , T.B. Fomby and J.Y. Park (eds.), Essays in Honor of Peter C. B. Phillips (*Advances in Econometrics*, Vol. 33), Emerald Group Publishing Limited, 425-490.

[30] Kiviet, Jan F. (2017). Discriminating between (In)valid External Instruments and (In)valid Exclusion Restrictions, *Journal of Econometric Methods* 6, 1-9.

[31] Kiviet, Jan F. and Milan Pleus (2017). The Performance of Tests on Endogeneity of Subsets of Explanatory Variables Scanned by Simulation, *Econometrics and Statistics* 2, 1-21.

[32] Knack, Stephen and Philip Keefer (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation, *Quarterly Journal of Economics* 112, 1252-1288.

[33] Leamer, Edward E. (1978). *Specification Searches: Ad Hoc Inference with Non-Experimental Data*, New York, NY, John Wiley and Sons.

[34] Lucas, Robert E. (1988). On the Mechanics of Economic Development, *Journal of Monetary Economics* 22, 3-42.

[35] Murray, Michael P. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments, *Journal of Economic Perspectives* 20, 111-132.

[36] Neumark, David, Junfu Zhang and Stephen Ciccarella (2008). The Effects of Wal-Mart on Local Labor Markets, *Journal of Urban Economics* 63, 405-430.

[37] Putnam Robert D. (1993). *Making Democracy Work: Civic Tradition in Modern Italy*, Princeton University Press. Princeton.

[38] Rosenzweig, Mark R. and Kenneth I. Wolpin (2000). Natural "Natural Experiments" in Economics, *Journal of Economic Literature* 38, 827-874.

[39] Thomas Sellke , M.J. Bayarri and James O. Berger (2001). Calibration of p Values for Testing Precise Null Hypotheses, *The American Statistician* 55, 62-71.

[40] Spring, Eva and Volker Grossmann (2015). Does Bilateral Trust Across Countries Really Affect International Trade and Factor Mobility?, *Empirical Economics* 50, 103-136.

[41] Staiger, Douglas and James H. Stock (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica* 65, 557-586.

[42] Stock, James H., Jonathan H. Wright and Motohiro Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments, *Journal of Business & Economic Statistics* 20, 518-529.

[43] Stock, James H. and Motohiro Yogo (2005). Testing for Weak Instruments in Linear IV Regression. In: Donald W.K. Andrews and James H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* New York: Cambridge University Press, 80-108.

[44] Tabellini, Guido (2008). Institutions and Culture, *Journal of the European Economic Association* 6, 255-294.

[45] Tabellini, Guido (2010). Culture and Institutions: Economic Development in the Regions of Europe, *Journal of the European Economic Association* 8, 677-716.

[46] Temple, Jonathan (2001). Growth Effects of Education and Social Capital in the OECD Countries, *OECD Economic Studies* 33, 57-101.

[47] Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, 2nd edition.

[48] Zak, Paul J. and Stephen Knack (2001). Trust and Growth, *Economic Journal* 111, 295-321.

Online-Appendix: Two Prominent But Falsified IV Studies

An important strand of macro-development literature has focussed on the question which fundamental factors could matter for economic exchange and economic development. We have seen that identifying such factors empirically requires structural estimations. To substantiate the point further, this Online-Appendix demonstrates that the absence of structural modelling can easily lead to erroneous conclusions. We next discuss two widely-received studies that have indeed been falsified and discuss what we could learn from them.

A. Institutions and Economic Development

Acemoglu, Johnson and Robinson (2001; henceforth AJR) address the important question to which extent the observed variation in institutional quality across economies can explain the variation in their stage of economic development. The link from institutions to economic development is indeed well grounded in economic theory.[22] Their measure of institutional quality is a "risk of expropriation index", supposed to capture institutions that secure property rights. Identification of causal effects on economic development is very challenging, however. As AJR (p. 1369f.) correctly note, potentially, "economies that are different for a variety of reasons will differ both in their institutions and in their income per capita". They suggest that "mortality rates of soldiers, bishops, and sailors stationed in the colonies between the seventeenth and nineteenth centuries" could serve as excluded instruments in a 2SLS approach, in which the "risk of expropriation index" is regressed on settler mortality at the first stage.

AJR note that some scholars have already suggested a link between the disease environment and prosperity in previous work, but at the same time emphasize that they "are not aware of others who have pointed out the link between settler mortality and institutions" (AJR; p. 1372). This is potentially problematic, however, as it basically

[22] For instance, see Gradstein (2008) and the references therein. Moreover, AJR themselves extensively and convincingly discuss that regions heavily differ in institutional quality.

means that there is no theoretical work to motivate the validity of their excluded instrument. Unfortunately, they do not provide a rigorous reasoning on the issue either. Consequently, it remains unclear why settler mortality may affect their "risk of expropriation index", leaving doubts on the instrument relevance assumption. Moreover, as AJR point out themselves, if settler mortality were related to the general disease environment, which is not controlled for and may have an effect on per capita income independently of institutions, the exclusion restriction would be violated.

Unsurprisingly, their OLS estimates suggest that (contemporaneous) institutions matter for (contemporaneous) per capita income. More surprisingly, the IV estimate of the coefficient on the "risk of expropriation index" is even larger in absolute value than the OLS coefficient estimate. An omitted variable which is also related to institutional quality, however, would most likely have opposite impacts on the "risk of expropriation index" and per capita income. There also may be the possibility that rich countries can better afford high-quality institutions. In these cases, the absolute value of the OLS estimate of the coefficient on the "risk of expropriation index" would be biased upwards, not downwards. AJR are obviously aware of that problem and refer the reader to a quite speculative explanation why the absolute value of the IV estimate exceeds that of the OLS estimate: "measurement error in the institutions variables that creates attenuation bias is likely to be more important than reverse causality and omitted variables biases" (AJR; p. 1385).

However, it may rather be the case that their excluded instrument is simply not valid. Lacking rigorous theoretical foundation, the *ex ante* probability that settler mortality affects property rights protection may in fact be viewed as low, potentially creating a severe weak instrument problem. A recent paper by Albouy (2012) substantiates this view. He shows that the construction of the mortality rates used in AJR (and in at least 20 subsequently published studies) is generally problematic. He argues that for 36 of the 64 countries in the AJR sample, mortality rates are extrapolated by assuming in a generally unfounded way that other countries have similar disease environments than the countries for which data within the own borders is available. As Albouy (2012) points

out:

> "the data come primarily from European and American soldiers in the nineteenth century. In some countries, rates apply to soldiers at peace in barracks, while in others the rates apply to soldiers on campaign. As is well known, soldiers on campaign typically have higher mortality from disease. This causes problems as AJR uses rates campaigns more often in countries with greater expropriation risk and lower GDP, artificially favoring the article's hypothesis. In a few countries, the data include the peak mortality rates of African laborers, but these are not comparable with average soldier mortality rates. Controlling for the source of the mortality rates weakens the empirical relationship between expropriation risk and mortality rates substantially. Furthermore, if these controls are added and the conjectured data are removed, the relationship virtually disappears, suggesting that it is largely an artifact of the data's construction" (p. 3060).

What does that imply for the IV estimation actually provided by AJR? The estimated coefficient on the settler mortality variable they actually employ is significantly different from zero at the first stage (regressing institutional quality on their measure of settler mortality rates). However, according to our analysis in section 3, because of the low *ex ante* probability that settler mortality affects institutional quality, the probability that it determines per capita income is low nevertheless. In fact, with the corrected measure of settler mortality provided by Albouy (2012), the first stage results suggest that the instrument relevance assumption is violated.[23]

B. Cultural Proximity and International Trade

In AJR the first stage regression of the 2SLS estimation was problematic, rooted in the absence of a sufficient theoretical foundation for the excluded instrument and data

[23] Moreover, even if the measure of settler mortality in AJR were (partially) correlated with the endogenous regressor, because we do not know what it actually measures, assuming that the exclusion restriction holds is speculative.

construction errors. We now come to an example where the main economic hypothesis, tested at the second stage of a 2SLS estimation, may be viewed as unconvincing to begin with.

Guiso, Sapienza and Zingales (2009; henceforth GSZ) examine the question to which extent the volume of exports from a country S (source) to a country D (destination) depends on the average level of trust citizens in D have towards citizens in S ("DtS trust"). Although there is strong theoretical foundation that bilateral trust is important for some forms of economic exchange between individuals (see, e.g., Arrow, 1972; Coleman, 1994), we are not aware of any theory that rigorously argues why at the aggregate (country) level an averaged measure of bilateral trust between ordinary citizens may affect international trade relations. In fact, those trade relations are often linked to international, professional networks of exporting firms and to the reputation of multinational companies.

Regressing international trade flows on measures of average bilateral trust in a sample of European countries, GSZ come up with OLS coefficients on bilateral trust that indeed are not significantly different from zero. This could have been taken as evidence to refute the hypothesis of interest. For instance, if there were a reverse causality problem of the sort that better trade relations could raise trust, the OLS coefficient on bilateral trust would even be biased upwards.

Nevertheless, GSZ continue with an IV strategy in the hope of finding evidence for their hypothesized positive effect of bilateral trust on bilateral trade. Econometrically, the challenge is to find a variable that affects international trade flows only through its effect on bilateral trust. GSZ propose two candidate instrumental variables that could shape DtS trust. The first is based on the idea that average physical dissimilarities across inhabitants of different countries (called "somatic distance") determine bilateral trust. The second measures the likelihood that two individuals in different countries share the same religion (called "religious similarity"). Constructing measures of both and using them as excluded instruments, GSZ find that the IV coefficient on bilateral trust at the second stage estimation becomes five times larger than the OLS counterpart, being

highly significant. Moreover, first stage results suggest a high (partial) correlation of the instrumental variables excluded at the second stage.

Given the good theoretical reasons suggesting that the OLS coefficient on bilateral trust is even biased upwards, the IV results of GSZ should be met with skepticism. In fact, Fehr (2009) convincingly argued that religious similarity may be expected to affect international trade through other channels than bilateral trust, suggesting that the IV strategy proposed by GSZ is likely to be invalid by violating the instrument exogeneity assumption. As GSZ admit, "it is possible that – test of overidentifying restrictions notwithstanding – our instruments are not orthogonal to trade, but pick up a set of cultural, institutional, and legal connections that facilitate trade flows" (p. 1120). For them, however, this is still not a reason to conclude that their approach is unlikely to be able to provide useful answers to their research question. Instead, they insist that exactly in the case where the instrument exogeneity assumption is violated "our results suggest the importance of culture-specific factors in trade relationships" (GSZ, p. 1120).

However, testing the hypothesis that cultural factors affect international trade based on their reasoning would call for investigating whether somatic distance and religious similarity matter in the reduced form of the dependent variable (international trade) – an analysis left out by GSZ. Spring and Grossmann (2015; henceforth SG) aimed to fill this gap. GSZ "compute the somatic distance between two countries as the sum of the absolute value of the difference in each of these traits" (p. 1107). Since somatic distance can be measured in various ways, in addition to replicating the original somatic distance indicator in GSZ, SG construct seven alternative combinations of the traits hair color, cephalic index, height and skin color to measure physical dissimilarities in a single index, based on the same data source GSZ use (Biasutti, 1959). For instance, they propose to weight the regions of a country according to the population density to compute the average measure of a physical trait in a country (unlike GSZ who focussed on the "majority trait").

Employing the alternative somatic distance indicators as excluded instruments in a 2SLS approach and in reduced form regressions of international trade, all of them turn

out to be partially correlated with bilateral trust equally strong than the original one employed by GSZ at the first stage. However, when instrumenting bilateral trust with any of the seven alternative indicators, second stage results are very different to GSZ. Trust coefficients become insignificant whenever somatic distance is the sole excluded instrument and sometimes are even negative.[24] In an reduced form analysis of the outcome variable as derived in eq. (4), where bilateral international trade flows are regressed on religious similarity and one of the (highly correlated) indicators of somatic distance at a time, the coefficient on religious similarity is never statistically different from zero and the coefficient on somatic distance is significant only when using the replicated original indicator of GSZ. The findings of SG thus raise strong doubts on the hypothesized relationship between bilateral trust or cultural proximity between countries (apart from the commonality of language) and bilateral international trade at the aggregate level.

It is striking that the paper by GSZ has been prominently published and widely received despite a weak theoretical foundation of the main hypothesis, a missing reduced form analysis of the dependent variable, and a high likelihood of invalidity of the excluded instrumental variables. One lesson from the replication analysis of SG is that future research should report estimation results based on many, differently constructed excluded instruments. Requiring to test whether the hypothesis of interest is supported with alternatively constructed excluded instruments would generally raise credibility of conclusions based on IV estimators. This particularly applies if researchers have a large degree of freedom to construct excluded instruments or can fish them out of a large universe of potential instruments.

Additional References

Arrow, Kenneth J. (1972). Gifts and Exchanges. *Philosophy & Public Affairs* 1, 343-362.

Biasutti, R. (1959). *Le Razze e i Popoli della Terra*, Unione Tipogra co - Editrice Torinese, Turin, 3rd edition.

[24] In most regressions, SG do not exclude religious similiarity at the second stage estimations, in line with the reasoning of Fehr (2009) about the potential violation of the instrument exogeneity assumption.

Coleman, James S. (1994). *Foundations of Social Theory*, Harvard University Press, Harvard.

Fehr, Ernst (2009). On the Economics and Biology and Trust, *Journal of the European Economic Association* 7, 235-266.

Gradstein, Mark (2008). Institutional Traps and Economic Growth, *International Economic Review* 49, 1043-1066.