Article

Compulsory Schooling and Returns to Education: A Re-Examination

Sophie van Huellen * and **Duo Qin**

SOAS University of London, Thornhaugh Street, Russell Square, London WC1H 0XG, UK
* Correspondence: sv8@soas.ac.uk; Tel.: +44-(0)-20-7898-4543

check for
updates

Abstract: This paper re-examines the instrumental variable (IV) approach to estimating returns to education by use of compulsory school law (CSL) in the US. We show that the IV-approach amounts to a change in model specification by changing the causal status of the variable of interest. From this perspective, the IV-OLS (ordinary least square) choice becomes a model selection issue between non-nested models and is hence testable using cross validation methods. It also enables us to unravel several logic flaws in the conceptualisation of IV-based models. Using the causal chain model specification approach, we overcome these flaws by carefully distinguishing returns to education from the treatment effect of CSL. We find relatively robust estimates for the first effect, while estimates for the second effect are hindered by measurement errors in the CSL indicators. We find reassurance of our approach from fundamental theories in statistical learning.

Keywords: instrumental variables; randomisation; research design; average return to education

JEL Classification: C26; C52; I21; I26; J24

1. Introduction

Over the past century, compulsory school law (CSL) was introduced in virtually every middle and high-income country (Goldin 1998; Goldin and Katz 2007). Empirical investigations into the effect of the CSL on educational attainment and income were pioneered by Angrist and Krueger (1991). The authors used CSL indicators as instrumental variables (Ivs) to 'randomise' latent ability across educational attainment groups to correct for the presumed inconsistency or beyond-sample bias in the ordinary least square (OLS) estimator. The empirical strategy is now common practice in research on the average return to education (ARTE) and the paper has since entered the standard economics curriculum, as evident from its appearance in two popular textbooks by Angrist and Pischke (2009, 2015).

Despite the far-reaching influence of this strategy, the causal interpretation of the CSL-treated schooling coefficient remains contentious. This is reflected in two interlinked developments. First, the emergence of IV estimates that vary significantly with the choice of instruments. Angrist and Krueger (1991), who approximate CSL with quarter of birth dummies, find that the IV estimates are not statistically different from estimates obtained via OLS.[1] Acemoglu and Angrist (2001) and Stephens and Yang (2014) replicate the research design by Angrist and Krueger (1991) with alternative CSL indicators based on labour law and find IV estimates which, although significantly different from OLS estimates, are insignificant or negative.[2] Second, a shift in the interpretation of the CSL instrumentalised returns

---

1   E.g., column (5) versus (6) in Table 4, (7) versus (8) in Table 5, and (1) versus (2) and (5) versus (6) in Table 6 in Angrist and Krueger (1991). Hoogerheide and van Dijk (2006, Table 5) and in Harmon et al. (2003, sec. 5).
2   See Angrist and Pischke (2015, Table 6.3) for a summary.

to schooling coefficient despite identical model choice. Angrist and Krueger (1991) interpret their results as consistent estimates of the ARTE, whereas Stephens and Yang (Stephens and Yang 2014, p. 1789) interpret their IV estimates as the effect of an additional year of education obtained due to CSL on income.

The first development prompts the question of how to select one consistent IV estimate among a multitude of IV choices. The second development prompts questions over the causal meaning of the IV estimates. The literature has responded to these questions by declaring certain instruments as inadequate, e.g., see Angrist and Pischke (2015, p. 227) for the above cases and Stock et al. (2002) and Kolesár et al. (2015) more generally, and by pointing at sample heterogeneity in the CSL effect, see Stephens and Yang (2014), Angrist et al. (1996), and Angrist and Imbens (1995) more generally.[3] However, the credibility of this empirical strategy is still disputed methodologically; see Deaton (2009) and Deaton and Cartwright (2018).

In this paper, we approach and analyse the contention from a different perspective. Drawing on fundamental concepts and theories from statistical learning, we argue that what is commonly described as a choice of consistent estimator is a choice of causal model design, whereby model choice has far more substantial implications for the consistency criterion than estimator choice. Further, a change in causal model design implies a change in the key causal variable, leading to a change in causal meaning of coefficient estimates. From this perspective, we can provide clarification regarding the questions raised and hopefully settle the methodological dispute. We demonstrate our arguments by replication and re-examination of two seminal studies by Angrist and Krueger (1991) (AK hereafter) and Stephens and Yang (2014) (SY hereafter).[4]

The insights gained from this new perspective are a consequence of two observations. First, the essence of the IV approach is the modification of a presumed endogenous causal variable, whereby the causal variable is substituted by regressors produced from non-uniquely and non-causally specified, and non-optimally targeted regressions; see Qin (2015, 2018) for a more detailed methodological exposition. Empirical evaluation and selection of these generated regressors is hence a source of endless contention. Second, the theoretical proof of IV estimator consistency rests on the presumption that the associated model specification is globally valid. This presumption is unlikely to hold in practice, as revealed by the out-of-sample error decomposition, known as the bias-variance tradeoff in the statistical learning literature. Analysis of this decomposition points to model bias rather than estimator bias as the primary source of inferential bias. Further, the presumption rules out any form of empirical model selection, including the choice between different instruments. This presumption is hence in conflict with the practical application of the IV approach.

Approaching the issue of modelling ARTE from this new perspective in Section 2, we show that the use of IV estimators amounts to making, albeit implicitly, the presumption of the education variable being an invalid conditional variable, thereby changing the causal model specification. Conceptualising the choice of the IV versus the OLS as one of causal model choice between non-nested model alternatives, this presumption can be explicitly specified into testable hypotheses. Moreover, the conceptualisation reveals the need to clarify the causal role of the CSL instruments. The causal chain representation method by Cox and Wermuth (2004) is applied to unravel the shift in interpretation of causal parameter estimates. While promiscuous in the IV approach, the chain representation makes possible the clear separation of two types of income effects: The ARTE with a possible moderation effect of CSL and the

---

[3] This argument is related to the programme evaluation modelling literature where the treatment variable, a dummy, is endogenised; e.g., Harmon et al. (2003) and Ludwig et al. (2012, 2013). The average treatment effect (ATE) estimate becomes a local ATE (LATE) estimate confined to the complier group if the instrument's effect is heterogeneous, e.g., see Angrist and Pischke (2009, chp. 4); Heckman and Urzua (2009); Deaton (2009); and Imbens (2010). However, this discussion is virtually irrelevant here as the treatment variable, i.e., CSL, has not been considered as endogenous in either studies.

[4] The two data sets used in these studies are both created from the 1980 US census but with different indicators and choice of control variables. The data used by SY is an extended version of the data and indicators used by Acemoglu and Angrist (2001)Acemoglu and Angrist

average treatment effect (ATE) of the CSL via schooling. The separation further enables us to assess risk of bias, i.e., omitted variable bias (OVB), measurement error, and selection bias, at the level of individual causal parameters.

In Section 3, we find no evidence of convergence as a necessary condition for consistency of the IV models, regardless the choice of instruments by k-fold cross validation (CV). CV is an essential tool for out-of-sample comparison of model generalisability, stability, and consistency in statistical learning. Further, decomposition of the two income effects, ARTE and the ATE of CSL via schooling, in Section 4 reveals that firstly, relatively robust ARTE estimates across cohorts and data sets can be obtained when carefully choosing covariates. Secondly, the estimated ATE effects of CSL and the CSL moderation effects on schooling are undermined by considerable measurement-error problems in the CSL indicators provided in AK and SY. Specifically, by careful choice of covariates, we find a virtually invariant and empirically consistent ARTE estimate of 0.06, and a smaller ATE of the CSL estimates between 1–5% if using labour law indicators and 0.2–0.9% if using quarter of birth indicators. It should be noted, however, that the empirical analysis is limited by the available covariates and instruments provided by AK and SY.

The empirical results in Section 4 show us how a causally explicit model design through statistical data learning enables us to clearly separate, empirically and conceptually, the causal meaning of parameter estimates and to assess the risk of inferential bias at the level of individual parameters. Methodological implications of these findings are extended in Section 5. Angrist and Pischke (2015, p. 227) discard the Acemoglu and Angrist (2001) study as 'a failed research design' and ascribe the failure to the choice of inappropriate CSL indicators. While we also find shortcomings in the CSL instruments, we delve deeper into the failure to reveal its root in equivocal causal model modifications by choosing the IV-based modelling approach. This choice virtually prevents direct and careful translation of causal postulates of interest into data-consistent conditional relationships. Although being constrained by the data sets provided in AK and SY, our re-examination of the CSL case clearly shows the importance of empirical model design and selection over estimator choice.

2. Model Specification of Schooling Effects Under CSL Treatment

The main objective of both AK and SY is to obtain consistent estimates of the effect of education on income, known as the ARTE. They reject, as inconsistent, the OLS in favour of the IV estimator. In contrast to previous literature, we transpose the OLS versus IV estimator choice into a choice of non-nested conditional models. This transposition leads us to re-evaluate the consistency claim underlying the choice of the IV approach and helps us to disentangle the seemingly conflicting causal interpretations presented in AK and SY. To facilitate the task, we adopt the subscript-based parametric notational methods used by Cox and Wermuth (2004) to highlight the consequence of different causal specifications on the parameters of regressors.

Denote education by s, and income by y, the OLS-based approach of estimating ARTE amounts to proposing the following simple regression model:

$$y = \alpha_y + \beta_{ys}s + \eta_y \qquad (1)$$

(1) is perceived as an invalid conditional model by both AK and SY on the presumption that $cov(s\eta) \neq 0$. The presumption is based mainly on the argument that (1) suffers from omitted variable bias (OVB), i.e., η_y contains variables which are not directly observable but collinear with s, such as aptitude. Their remedy is to utilise the CSL as a key instrument to block this bias. Specifically, the following regression is used to generate s^L, the fitted response from (2):

$$s = \pi_{sL.I_j}L + I'_j\gamma_{sI_j.L} + e_s \Rightarrow s^L, \qquad (2)$$

where L represents the CSL and I_j a vector of other Ivs. (2) is commonly referred to as the first stage of the two-stage least square (2SLS) estimator, to facilitate the following second-stage equation:

$$y = \alpha_y + \beta_{ys^L} s^L + \eta_y^L. \tag{3}$$

From the perspective of model specification, (1) and (3) are de facto non-nested models. The necessary condition for having statistically significant $\beta_{ys} \neq \beta_{ys^L}$ is to generate s^L, such that $s^L s$.[5] In general, since no unique set of Ivs exist for (2) in practice, it is impossible to settle a priori on one unanimously agreed definition of s^L.[6] That implies that (3) should be seen as representing a multitude of non-nested models. Modellers are compelled to go through a model selection process, albeit implicitly through experimenting with various IV sets, as seen in both the AK and SY cases. One drawback of this implicit practice is the lack of model selection rules for guidance.

Once the task is recognised as one of model selection rather than estimator selection, out-of-sample cross validation (CV) methods, which are widely used in statistical learning, emerge as a useful toolbox to evaluate beyond-sample inferential bias. According to statistical learning theory, model selection is targeted at structural risk minimisation over a given hypothesis space that spans over the competing model specifications. A model is selected against its alternatives based on the interlinked criteria of generalisability (or predictivity), stability, and consistency, whereby Mukherjee et al. (2006) show that stability is equivalent to empirical consistency. CV methods are designed to assess predictivity and consistency by splitting the sample into k-folds, with k-1 folds being used to train the model and the kth fold to test the model. The competing model specifications can hence be evaluated by comparison of the relative mean squared error (MSE) in a k-fold CV, e.g., see Arlot and Celisse (2010), Shalev-Shwartz et al. (2010) and Zhang and Yang (2015).[7]

At the core of CV methods is the analysis of MSE through its decomposition into bias and variance, and the demonstration of the tradeoff between the two components. In particular, the analysis identifies model bias as the primary source of inferential bias, i.e., the bias component in the out-of-sample or the testing sample errors. Another fundamental insight from statistical learning is the recognition that theoretical models, i.e., formal constructs of prior knowledge, are the source of inductive bias. Hence, in the quest for structural risk minimisation, major attention is paid to the minimisation of inductive bias in model selection and model design, see e.g., Shalev-Shwartz and Ben-David (2014, Part I). In light of these fundamental theories, we see the need, in addition to the application of CV, of scrutinising carefully the process of how schooling effects on income under the CSL treatment are formalised into (3). Especially, whether the various contextual reasons supporting its formalisation, such as OVB and related measurement errors as well as selection bias, can justify the rejection of (1).

Since the CSL effect on income via schooling is a sequential event, this can be represented by a reduction of the following recursive factorisation of the joint density, $f(y, s, L)$:

$$f(y, s, L) = f(y|s, L)f(s|L)f(L) = f(y|s, L)f(s|L), \tag{4}$$

since $f(L) = 1$ when retrospective cross-section data samples are used. When L is assumed to act as a rule of intervention, namely $y \perp L | s$, the conditional density in (4) can be further factorised:

$$f(y, s|L) = f(y|s)f(s|L). \tag{5}$$

On the basis of (5), we can express the sequential nature of the ATE of L on y via s by the conditional expectation, $E(y, s|L) = E(y|s)E(s|L)$. In a linear model setting, this expectation decomposition leads to the following chain model representation, see Cox and Wermuth (2004):

$$\begin{aligned} y &= \alpha_y + \beta_{ys} s + \eta_y \\ s &= \alpha_s + \beta_{sL} L + \eta_s \end{aligned} \tag{6}$$

5　Notice that this requirement imposes a non-optimal prediction constraint on (2), in that the specification of this regression must *avoid* explaining the response variable as accurately as possible.

6　See Qin (2015, 2018) for a more detailed analysis of the causal model modifying roles of the IV approach.

It should be noted that (6) differs from (2) + (3) in two substantial ways. First, β_{ys} still embodies ARTE in (6). Second, the ATE of L on y, denoted by β_{yL}, is derivable from $\beta_{yL} = \beta_{ys}\beta_{sL}$, whereas there lacks a clear parametric representation of this effect in the IV model. Although $\pi_{sL.I_j}$ in (2) can be interpreted as the ATE on s, this parameter cannot be used in conjunction with β_{ys^L} in (3) to identify the ATE on y.

An arguably useful tool to highlight these differences is the directed acyclic graph (DAG), see Cox and Wermuth (1996) and Wermuth and Cox (2011). The left panel of Figure 1 is a DAG of (6). It shows us that, when ATE, i.e., the effect of L forms the focal causal interest, s takes the role of an intermediate variable or a mediator, but when ARTE is the focal interest, L takes the role of a moderator exclusively for s. The second arrow segment, $L \rightarrow s$, can be ignored in the latter case, i.e., the case when (6) is reduced into (1). The middle panel is a DAG of (2) + (3).[8] This IV-based model is focused on the first arrow segment, since its objective is to reject (1). Hence, the possibility of a causal chain extension is blocked, and L is used to target at producing $s^L s$, making ATE of L on y unidentifiable—but neither is ARTE identifiable because s has been significantly modified. Therefore, the definition of β_{ys^L} needs to be modified.
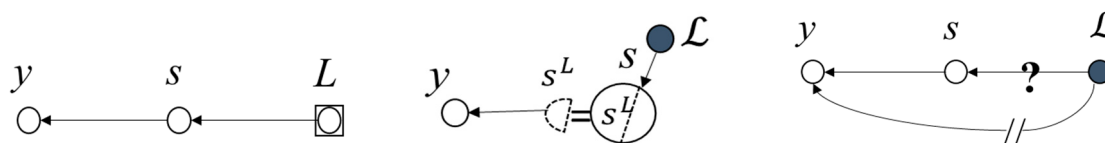

Figure 1. Directed acyclic graphs (DAGs) of returns to schooling under the compulsory school law (CSL) treatment. Notes: y denotes earnings; s, schooling; L, the CSL; and $\mathcal{L}$, its observable indicator. A node inside a square indicates a latent variable, and a solid node denotes a dummy/binary variable. Dotted lines indicate non-uniqueness; dissimilarity of s^L from s is shown by a semicircle; the 'identity' sign differentiates the first stage of the 2SLS, (2).

Now, we are in the position of examining the contextual reasons underlying (3) to find out whether the IV-induced modification helps resolve the problems that the approach is intended. Although OVB is stated as the primary problem by AK, it is further compounded, in their justification for the IV route, with two problems—measurement error and selection bias. In view of the current modelling purpose, the concern over measurement error in s is unwarranted because ARTE is not ARTA (average returns to aptitude).[9] In other words, measurement error is irrelevant in (1) unless we change its prior stance to explicitly specify s as an imperfect indicator of the latent variable, 'aptitude'. However, measurement error can provide β_{ys^L} in (3) with a plausible interpretation differently from that of β_{ys}. However, this interpretation would undermine the basic IV-based claim of β_{ys^L} being the consistent estimator of ARTE with respect to s, and openly recognise (1) and (3) as two different models, with (3) effectively yielding ARTA. As for selection bias, the argument extends to the situation where CSL treatment could alter the population composition of educated workers, as compared to that of the pre-treatment population, e.g., through a diluted concentration level of 'aptitude' (see Angrist and Pischke 2009, chp. 4). Consequently, the post-treatment schooling effect becomes significantly different from the pre-treatment one due to a change in level of 'aptitude' for different years of schooling post-treatment. Two problems hinder this argument. First, there lacks a credible way to verify that a compositional shift, if it has occurred, is adequately reflected by s^L generated via (2). From the perspective of retrospective cross-section data, empirical assessment of the possibility of such a shift entails disaggregation. Specifically, we need

---

[8]　Unfortunately, DAGs in several existing publications have misrepresented the IV approach as one of causal chain extension, e.g., Figure 7.8 in Pearl (2009) and Figure 6 in Abadie and Cattaneo (2018).

[9]　The inapplicability of the measurement errors-based arguments in the present context can also be seen from the fact that almost no signs of expected OLS attenuations caused by measurement error concerns can be found in AK or SY, namely that the OLS estimates should be statistically insignificant and smaller in magnitude than the IV estimates, e.g., Durbin (1954).

to carefully divide the available samples into two parts—an *L*-treated part versus a CSL unaffected part—so as to investigate whether there exists a parametric difference: $\beta_{ys_L.Z} \neq \beta_{y\widetilde{s_L}.Z}$, where s_L denotes schooling of the *L*-treated part, and $\widetilde{s_L}$ the treatment unaffected part.[10] Even if the inequality is supported by data, the evidence alone is insufficient for rejecting *s* as a valid conditional variable for *y* at the aggregate level, e.g., see Engle et al. (1983) and also Qin et al. (2019). Second, the argument assumes a role of *L* in conflict with its role in the IV treatment of OVB—that the instrument must be unrelated to the omitted variable under the suspicion of causing OVB.

The above analysis not only casts doubt over the explanatory capacity of the IV-based model (3), but also draws our attention to the need to clarify the expected role of *L* in accordance to our modelling purposes. Clearly, if ATE forms part of our inferential interest, we should not reduce model (6) to (3). Let us turn to this treatment effect. Model (6) tells us that $\beta_{yL} = \beta_{ys}\beta_{sL} \neq \beta_{ys}$ in general unless $\beta_{sL} = 1$ can be verified, which is highly unlikely in view of available findings, e.g., see Goldin and Katz (2011). Hence, we should expect that $\beta_{yL} \ll \beta_{ys}$. However, if ATE is the only parameter of our interest, the chain route of (6) appears a long way round, because β_{yL} can be estimated directly from:

$$y = a_y + \beta_{yL}L + \epsilon_y. \tag{7}$$

Unfortunately, this direct route is unfeasible in the samples used by AK and SY because *L*, a notional variable for CSL, is latent and approximated by various observable indicators, $\mathcal{L}$. Consequently, measurement errors are likely to result in $\beta_{y\mathcal{L}} \neq \beta_{ys}\beta_{s\mathcal{L}}$, when $\mathcal{L}$ is used in (7) instead of *L*. For instance, SY have identified this kind of defectiveness of CSL indicators, due to their entanglement with regional factors and other controls. On the other hand, a particular case of $\beta_{y\mathcal{L}} \neq \beta_{ys}\beta_{s\mathcal{L}}$ signals its associate $\mathcal{L}$ being a defective indicator, as it fails to embody the assumed rule of intervention. This failure can be identified via checking $\beta_{y\mathcal{L}.s} \neq 0$ of the following regression:

$$y = \alpha_y + \beta_{ys.\mathcal{L}}s + \beta_{y\mathcal{L}.s}\mathcal{L} + \varepsilon_y. \tag{8}$$

In other words, a test of $\beta_{y\mathcal{L}.s} = 0$ using (8) can be exploited as an additional criterion for the purpose of $\mathcal{L}$ selection; see Zhang et al. (2017) for implications of measurement error in estimating causal chain models. A DAG illustration of this situation is given in the right panel of Figure 1.

The advantage of the chain route becomes even more evident when the presence of control variables, denoted by *Z*, is taken into consideration. Although *Z* is chosen primarily from consideration of $cov(sZ) \neq 0$, some variables in *Z* are likely to be correlated with $\mathcal{L}$, such as age and regional dummies in the two data sets by AK and SY. The DAGs with *Z* included are shown in Figure 2. The potential correlation would complicate the estimation of ATE. Extend (6) by *Z*:

$$\begin{aligned} y &= \alpha_y + \beta_{ys.Z}s + Z'\boldsymbol{\beta}_{yZ.s} + \varepsilon_y \\ s &= \alpha_s + \beta_{sL}L + \varepsilon_s \\ Z &= \boldsymbol{\alpha}_Z + L\boldsymbol{\beta}_{ZL} + \boldsymbol{\varepsilon}_Z \end{aligned} \tag{9}$$

The corresponding chain representation of the ATE becomes decomposed into two parts:

$$\beta_{yL} = \beta_{ys.Z}\beta_{sL} + \boldsymbol{\beta}'_{yZ.s}\boldsymbol{\beta}_{ZL} = \beta_{yL_s} + \beta_{yL_Z}. \tag{10}$$

Now, only the first component, β_{yL_s}, in (10) corresponds to the ATE of *L* via *s*. Model (7) is not fit for estimating this parameter.

---

[10]　We have empirically evaluated the hypothesis of a structural shift. Results are detailed in Appendix B. We find no supporting evidence of such shift.

Figure 2. DAGs augmented with Z. Notes: See the notes in Figure 1 for the definitions of the various symbols.

3. Evaluation of Model Consistency

Section 2 has shown that the IV approach amounts to a model re-specification by replacement of s with s^L as the valid conditional variable, thereby altering the causal interpretation of the coefficient estimates. This re-specification is based on the premise of inconsistency of the OLS model specification relative to its IV counterpart. By exposing the IV approach as a model re-specification, the estimator choice is transposed into one of non-nested model selection, which is testable by use of CV.

In the following, we first replicate results presented by AK and SY, while focusing mainly on SY, to identify conditions under which instrumental validity is achieved and then, by use of CV, reassess these results against the criteria of generalisability and consistency. Since the CSL is latent, it is approximated by observable indicators, $\mathcal{L}$. Quarterly birth dummies are chosen by AK ($\mathcal{L}_{AK}$).[11] SY, with reference to Acemoglu and Angrist (2001), propose two alternative indicators based on state school and labour law. These indicators capture required years of schooling ($\mathcal{L}_{SY1}$) and compulsory attendance ($\mathcal{L}_{SY2}$).[12]

Let us inspect the replicate of SY's results (see Figure 3). The IV-based model specifications appear to lack empirical consistency and robustness relative to their OLS counterpart. β^L fails to show convergence and standard errors remain large as the sample size increases. Although these findings are common in the literature, their implications are rarely discussed; see Deaton and Cartwright (2018).

Different choices of CSL indicators for generating different s^L result in considerable alteration of the estimation results in SY, as compared to AK. Only in SY, the choice of indicators leads to an apparent success in finding $\beta^L \neq \beta$. Further scrutiny through replication of SY's Tables 1 and A2 suggests that their CSL indicators are largely invalid instruments. Column (1) in T1B of our Table 1 is the only exception, with no rejection of Sargan's null of valid overidentifying restrictions and rejection of Hausman's null of OLS estimator consistency relative to IV. Although the validity of instruments is not rejected for column (2) in T1A of Table 1, the IV estimates remain insignificant.

In contrast to s, s^L seems to strongly correlate with covariates such as interaction terms that allow for regional differences in year of birth effects. The inclusion of these interaction terms leads to large changes in β^L, whereas β remains virtually invariant; see Figure 3 and columns (2) and (4) of T1A and T1B in Table 1. At the same time, the inclusion of interaction terms invalidates the claim of endogeneity if using $\mathcal{L}_{SY2}$ indicators and leads to an insignificant β^L estimate if using $\mathcal{L}_{SY1}$ indicators. The sensitivity of IV estimates to regional factors has already been pointed out by SY and reiterated by Hoogerheide and van Dijk (2006, Table 5).[13] This raises the question of whether $\mathcal{L}_{SY}$ solely represent the CSL treatment; a potential case of measurement error in these indicators.

---

[11] The indicator choice is based on the insight that the CSL requires a minimum age which must be reached before students can drop out of school. Those born in the first quarters of the year reach this age sooner than those born in later quarters and hence are less constrained by the law than their peers. Accordingly, AK define three birth dummies for those born in the first ($\mathcal{L}_{AK}^1$), second ($\mathcal{L}_{AK}^2$), and third ($\mathcal{L}_{AK}^3$) quarter of the year; see also Angrist and Krueger (1992).

[12] As in AK, the indicators compose of three dummies. $\mathcal{L}_{SY1}^1$, $\mathcal{L}_{SY1}^2$, and $\mathcal{L}_{SY1}^3$ capture those with minimum of 7 or below, 8, and 9 or above required years of schooling and $\mathcal{L}_{SY2}^1$, $\mathcal{L}_{SY2}^2$, and $\mathcal{L}_{SY2}^3$ capture those with 8 or below, 9, and 10 or above years of compulsory school attendance. See SY for a detailed definition of the indicators.

[13] CSL indicators based on quarter of birth dummies face similar problems, and Bound and Jaeger (2000) and Carneiro and Carneiro and Heckman (2002) show an entanglement of indicators with social status.
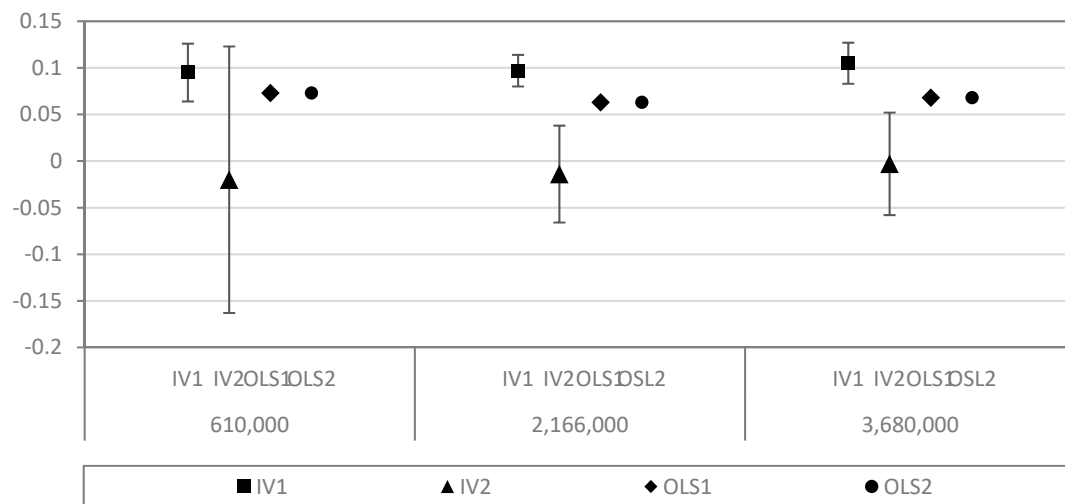
Figure 3. Ordinary least square (OLS) and instrumental variable (IV) estimator consistency. Notes: IV1 and OLS1 are IV and OLS estimates without regional control variables, and IV2 and OLS2 are IV and OLS estimates with regional control variables included. The x-axis provides the sample size and the y-axes coefficient values. The bars indicate the 95% confidence interval. Source: SY, Table 1.

Table 1. Sargan and Hausman test for instruments used by SY.

	T1A ($\mathcal{L}_{SY1}$)				T1B ($\mathcal{L}_{SY2}$)			
White Males	**Aged 40–49**		**Aged 25–54**		**Aged 40–49**		**Aged 25–54**	
	(1)	**(2)**	**(3)**	**(4)**	**(1)**	**(2)**	**(3)**	**(4)**
β (OLS) [a]	0.073 **	0.073 **	0.063 **	0.063 **	0.073 **	0.073 **	0.063 **	0.063 **
β^L (2SLS) [a]	0.095 **	−0.020	0.097 **	−0.014	0.142 **	0.092 **	0.140 **	0.086 **
Tests:								
Sargan [b]	0.99	4.65	17.99	7.51	0.64	0.83	12.75	17.57
(*p*-value)	(0.6088)	(0.0977)	(0.0001)	(0.0234)	(0.7271)	(0.6589)	(0.0017)	(0.0002)
Hausman	3.80	9.67	43.24	36.32	16.33	0.53	150.29	3.28
(*p*-value)	(0.0512)	(0.0019)	(0.0000)	(0.0000)	(0.0001)	(0.4671)	(0.0000)	(0.0701)
Fixed effects:								
State of birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year of birth	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region x Yob	No	Yes	No	Yes	No	Yes	No	Yes
Additional controls:	None	None	Age quartic, census year	Age quartic, census year	None	None	Age quartic, census year	Age quartic, census year

Notes: Sargan and Hausman added through replication. Source: SY Tables 1 and A2. [a] Robust and cluster adjusted standard errors are used. [b] Wooldridge's extension of Sargan's test of overidentifying restrictions is performed. ** Significant at the 1% level.

The insignificance and empirical inconsistency of β^L, identified in Figure 3 and Table 1, could also be caused by a negligible share of 'compliers' in the full sample; a point made by Oreopoulos (2006) in the context of the CSL effect when using minimum years of schooling indicators, and also mentioned by SY as a possible explanation for their results. We hence investigate whether instrumental validity can be achieved when focusing on a sub-sample with a high complier share. We follow AK's lead and divide the sample by those obtaining 12 years of schooling (school) and those who obtain more than 12 years of schooling (higher). The former sub-sample has a high share of compliers, while the latter sub-sample comprises mainly always takers. Results are reported and discussed in Appendix A. We do not find stronger evidence for instrument validity but make two interesting observations. Firstly, most IV estimates turn insignificant for the 'higher' sub-sample, confirming the high share of always takers. Secondly, OLS estimates reveal shrinking ARTE for those with more years of schooling, especially for the 1940–1949 born cohort. This cohort entered the labour market in the early 1980s in the middle of a

recession, potentially explaining the low returns to higher education. This effect is concealed in the IV models, given the localness of the CLS instruments.

We now turn to CV to formally compare the two non-nested models (1) versus (3). Figure 4 shows that the OLS-based models clearly outperform the IV-based models in generalisability, stability, and consistency, even though the CV experiment presented here does not adjust for degrees of freedom.[14] Remarkably, results for AK are close to—or even worse than for—SY, despite the finding of $\beta \approx \beta^L$ in AK.
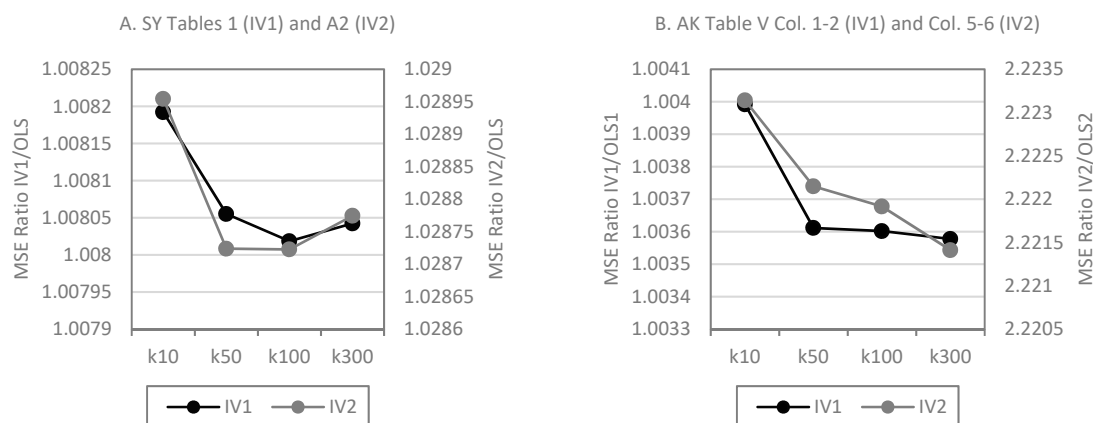

Figure 4. Ratio of average mean square cross validation (CV) error of IV to OLS with Increasing K. Notes: Mean squared error (MSE) is the average of 10 repetitions of the k-fold CV. The curves represent the ratio of the average MSE of the IV model and the OLS counterpart. A value greater than 1 indicates a smaller MSE for OLS than for IV.
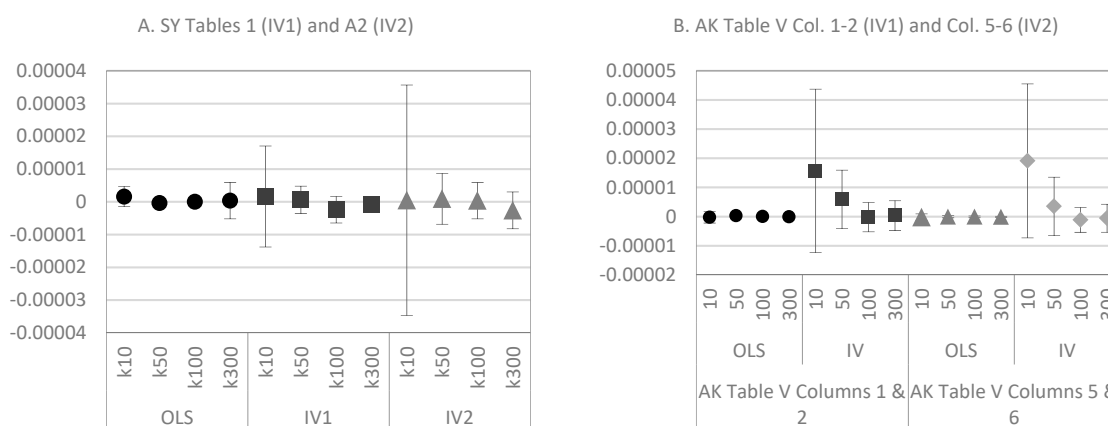

Figure 5. Average CV bias with increasing k. Notes: The CV bias is the average of 10 repetitions of the k-fold CV. Bars on the estimation bias indicate one standard deviation over the 10 repetitions. Numbers of folds shown on the x-axis.

As expected, the MSE decreases as the training sample increases, that is, with increasing k, for both models. However, the IV-based model shows no sign of convergence as training samples grow. When decomposing the MSE into test bias and variance, we find little evidence of asymptotic bias in the OLS estimates; see Figure 5. For the AK case, the IV bias is larger at smaller k and decreases towards no bias at larger k. Given the small bias in general, the large difference in the MSE between the two models clearly stems from a greater variance of the IV model specification, putting the consistency

---

[14] The IV approach uses up more degrees of freedom than the OLS counterpart due to the first stage. Therefore, the MSE of the IV model specification understates the error when compared to the OLS counterpart.

claim of IV into question. While our findings are specific to the CSL case and the chosen instruments, results by Young (2017), who re-evaluates 1359 published IV regressions, suggest that the conclusion drawn from the CV exercise are the norm rather than an exception.

Overall, experimenting with the model design in SY and AK, we find, contrary to what is expected, that the OLS-based model outperforms the IV-based ones in terms of generalisability, stability, and consistency, regardless the choice of CSL indicators.

4. Different Income Effects

Section 3 has provided us with no evidence for *s* being in invalid causal variable and we now probe into the causal role of *L*. In Section 2, we could distinguish between two income effects: (a) The ARTE effect, $\beta_{ys.Z}$, and (b) the CSL effect or ATE of the CSL via schooling, β_{yL_s} as specified in (9).

4.1. Estimating ARTE: $\beta_{ys.Z}$

The presentation of varying OLS-based ARTE estimates by AK and SY, despite the use of almost identical samples, indicates problems in the choice of appropriate covariates. Therefore, we proceed with the question of how to specify *Z* in order to find an empirically adequate specification of (9), which is as parsimonious as possible and can also align the ARTE estimates by AK and SY data, respectively. This is achieved through, firstly, unification of the coding of the education variable and secondly, a parsimonious model specification.

Towards a unification of the education variable, the AK education variable is capped at 17 years to resemble the SY education variable. The unification is found to play a vital role in aligning the ARTE estimates across the two data sets.[15] We rely on AK's division between those born in the 1930s and 1940s, respectively, using observations from the 1980 census. Towards a more parsimonious model, year of birth dummies included by both AK and SY are replaced with quadratic age (*age2*).[16] Regional dummies for individual states are replaced by a single variable distinguishing between four regions for SY and nine regions for AK data (*region*). Considering a possible regional effect on school quality, variables capturing school quality (*pupilt*, *term*, *reltwage*) suggested by Card and Krueger (1992a, 1992b) are used by SY and included in our model as well.

Earlier sub-sample experiments, reported in Appendix A, reveal variation in the ARTE estimates with the level of education. The variation reflects 'sheepskin effects', which are well documented phenomena in the literature[17] and clearly discernible in the AK and SY data; see Figure A1 and Table A2, Appendix A. A binary variable (*uni*) is thus added as a classifier for those who obtained a university degree (15 or more years of schooling).

The key results of this model search are reported in Table 2, alongside those from the 'Original' models by SY and AK. We refer to our more parsimonious model specifications as 'Alternative' in the table. A closer alignment of return to schooling estimates across data sets is achieved with the 'Alternative' model specification, which outperforms the 'Original' specification in terms of model fit by a margin. OLS estimates point to a relatively constant $\beta_{ys.Z}$ of about 0.06 across data sets and cohorts, and our ARTE estimates are roughly in line with findings by Acemoglu and Angrist (2001), who report estimates of 0.061 and 0.075 respectively.

---

[15] Ideally, we would use uncapped schooling variables, but the transformation in the SY schooling variable is irreversible.

[16] Coefficients on year of birth dummies are found to decline with years, revealing non-linearity. These patterns can be almost perfectly replicated with a quadratic age variable. See also Murphy and Welch (1992) for the non-linear relationship between experience and wage earnings.

[17] See, for instance Angrist (1995); Murphy and Welch (1992); Card (2001); Trostel (2005); and Clark and Martorell (2014). This shift in the population education composition also explains the finding by Goldin and Katz (2000).

Table 2. Parsimonious specification of (9).

	Original				Alternative			
	SY [a]		AK		SY [a]		AK	
	1930–1939	1940–1949	1930–1939	1940–1949	1930–1939	1940–1949	1930–1939	1940–1949
$\beta_{ys,Z}$	0.0751 **	0.0622 **	0.0630 **	0.0519 **	0.0600 **	0.0643 **	0.0576 **	0.0648 **
[95% CI]	[0.074–0.077]	[0.061–0.063]	[0.062–0.064]	[0.051–0.053]	[0.058–0.062]	[0.063–0.066]	[0.057–0.059]	[0.064–0.066]
AIC [b]	714,262.9	1,034,376	594,994.7	858,645.2	705,271.2	1,018,112	594,343.4	858,594.8
Adj.-R2	0.0119	−0.0232	0.1745	0.1354	0.1217	0.0968	0.1761	0.1355
Consist. [c]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Z [d]	age2 age3 age4 yob31-yob39/yob41-yob49 sob1-sob55		ageq, ageq2, race, married, smsa, neweng midatl, enocent, wnocent, soatl, esocent, wsocent, mt, year20-year 28		age2, mar, emp, jail, handcap, pupilt, term, reltwage, uni, region		age2, married, race, smsa, uni, region	

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. [a] 95% confidence interval based on cluster adjusted standard errors in SY data. [b] Akaike information criteria. [c] Entner et al. (2012) test for consistency. The row reports the correlation coefficient between ε_y in (9) and the residuals from the auxiliary regression, with a value close to 0 confirming consistency. Non-Gaussianity of the residuals was tested before and strongly supported by data. [d] See SY and AK for variable names. ** Significant at the 1% level.

Following the observations in Table 2, we note that the risk of OVB for $\beta_{ys.Z}$ comes from inadequately specified Z. Hence, we evaluate the choice of Z by use of a simple statistical test of consistency developed by Entner et al. (2012). Recalling the DAG in Figure 2, we can immediately see that in the presence of OVB, that is, missing covariates in Z, the residuals ε_y in (9) would be statistically dependent on s. Entner et al. (2012) exploit this insight by means of a simple two-step algorithm to test the consistency of $\beta_{ys.Z}$ against the risk of OVB. In a first step, the key conditional variable s is regressed on the set of covariates Z.[18] If residuals of this auxiliary regression are non-Gaussian—Gaussian residuals are a rarity in large cross-sectional data sets—it is tested for being statistically independent between ε_y from (9) and the error term of the auxiliary regression in a second step. If independence is confirmed, $\beta_{ys.Z}$ is consistent with regards to the choice of covariates Z. The test results are reported in the last row of Table 2. In all cases, consistency is strongly supported by the data.

4.2. Estimating the ATE of the CSL via Schooling: β_{yL_s}

Given the potential measurement error in CSL indicators identified by SY and briefly discussed in Section 3, we follow Section 2 and conduct two simple experiments to further test the appropriateness of the indicator choice before continuing with the estimation of β_{yL_s}. Since CSL is only binding for school leavers, we would expect the ATE to be insignificant or at least smaller for those with higher education than for those without. Following this reasoning, we estimate the middle equation of (9) using sub-sample groups by educational attainment with the expectation that $\hat{\beta}_{sL} \neq 0$ for *School* and $\hat{\beta}_{sL} = 0$ for *Higher*.

It is shown in Table 3 that, although $\hat{\beta}_{sL}$ tends to be larger for the *School* sub-sample than for the *Higher* sub-sample, none of the indicators confirms the hypothesis of $\hat{\beta}_{sL} = 0$ for *Higher*. Noticeably, the size of those $\hat{\beta}_{sL} \neq 0$ in the first cohort has almost doubled that of the second cohort in the case of SY indicators. This shift appears to reflect a general shift towards more years of education. As seen from Table A1 (Appendix A), the share of those attaining less or equal the minimum years of schooling is halved in the later cohort.

---

18   The auxiliary regression takes the form $s = \alpha_y + Z'\beta_{yZ.s} + \varepsilon_s^{aux}$. If ε_s^{aux} is non-Gaussian, statistical independence between ε_s^{aux} and ε_y confirms consistency of $\beta_{ys.Z}$ in (9).

Table 3. β_{sL} in (9) via sub-sampling on educational attainment.

	SY								AK			
	$\mathcal{L}_{SY1}$				$\mathcal{L}_{SY2}$				$\mathcal{L}_{AK}$			
1930–1939	School		Higher		School		Higher		School		Higher	
	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [b]	Coef.	t-stat [b]
$\beta_{s\mathcal{L}_1}$	0.38 *	2.50	0.06	0.69	0.18	1.84	−0.1 **	−4.33	−0.1 **	−8.74	0.02	1.17
$\beta_{s\mathcal{L}_2}$	0.35 *	2.49	0.07	0.91	−0.02	−0.23	−0.05 *	−1.98	−0.1 **	−8.43	0.05 **	3.10
$\beta_{s\mathcal{L}_3}$	0.18	1.25	0.06	0.75	0.39 **	3.99	−0.1 **	−4.27	−0.03 *	−2.39	−0.00	−0.04
	$\mathcal{L}_{SY1}$				$\mathcal{L}_{SY2}$				$\mathcal{L}_{AK}$			
1940–1949	School		Higher		School		Higher		School		Higher	
	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [a]	Coef.	t-stat [b]	Coef.	t-stat [b]
$\beta_{s\mathcal{L}_1}$	0.70 **	11.34	0.13 **	3.47	0.24 **	4.59	−0.05	−1.57	−0.1 **	−9.89	0.04 **	3.71
$\beta_{s\mathcal{L}_2}$	0.69 **	10.61	0.25 **	7.63	−0.03	−0.57	0.03	1.00	−0.1 **	−7.82	0.06 **	5.09
$\beta_{s\mathcal{L}_3}$	0.42 **	6.25	0.19 **	5.75	0.21 **	3.40	−0.06 *	−2.11	−0.02 *	−2.45	0.03 *	2.25

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. [a] Robust cluster adjusted standard errors. [b] Robust standard errors. ** Significant at the 1% level. * Significant at the 5% level.

In a second step, we test whether the rule of intervention $\beta_{y\mathcal{L}.Zs} = 0$ holds for the different CSL indicators by estimation of (8) with additional controls Z. In reference to earlier experiments, we conduct the test for the *School* sub-sample in addition to the full sample estimation. It is shown in Table 4 that the condition $\beta_{y\mathcal{L}.Zs} = 0$ is validated for SY's $\mathcal{L}_{SY1}$ indicator across cohorts and also for AK's $\mathcal{L}_{AK}$ indicator for the early born cohort. However, it is violated without exception if using $\mathcal{L}_{SY2}$ as CSL indicator. Where conditional independence is rejected in Table 4, we have also failed to confirm $\hat{\beta}_{sL} = 0$ for the *Higher* sub-sample in Table 3, and rejected instrument validity in Tables 1 and A2 Appendix A. In cases like this, we should be cautious with the estimate of β_{yL_s} via the chain representation of (10).

Table 4. Test for the rule of intervention $\beta_{y\mathcal{L}.Zs} = 0$ using (8) extended by Z.

	SY [a]								AK [b]			
	1930–1939				1940–1949				1930–1939		1940–1949	
	Full		School		Full		School		$\mathcal{L}_{AK}$		$\mathcal{L}_{AK}$	
	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	Full	School	Full	School
$\mathcal{L}1$	−0.001	0.041 **	0.007	0.041 **	0.014	0.041 **	0.007	0.046 **	−0.007 *	−0.008 *	−0.01 **	−0.005
	(0.0133)	(0.0081)	(0.0185)	(0.0100)	(0.0125)	(0.0099)	(0.0146)	(0.0106)	(0.0030)	(0.0039)	(0.0024)	(0.0035)
$\mathcal{L}2$	0.011	0.028 **	0.029	0.030 **	0.016	0.047 **	0.020	0.052 **	−0.004	−0.009 *	0.012 **	0.010 **
	(0.0139)	(0.0062)	(0.0189)	(0.0078)	(0.0126)	(0.0081)	(0.0142)	(0.0085)	(0.0030)	(0.0039)	(0.0024)	(0.0035)
$\mathcal{L}3$	0.021	0.060 **	0.033	0.063 **	0.033 **	0.078 **	0.031 *	0.082 **	0.001	−0.003	0.012 **	0.017 **
	(0.0144)	(0.0103)	(0.0196)	(0.0113)	(0.0122)	(0.0122)	(0.0140)	(0.0124)	(0.0030)	(0.0038)	(0.0023)	(0.0034)

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. Z as specified in 'Alternative' in Table 4. Standard errors reported in parentheses. [a] Robust cluster adjusted standard errors. [b] Robust standard errors. ** Significant at the 1% level. * Significant at the 5% level.

Table 5 provides β_{yL_s} estimated via (10). Where conditional independence is verified, the chain approximation yields significant ATE estimates that confirm our expectation of $\beta_{ys.Z} \gg \beta_{yL_s}$. The estimated ATE almost doubles for the later born cohort from 1–3 to 3–5% using $\mathcal{L}_{SY1}$ indicators. The ATE estimates using AK indicators are relatively constant across both sub-samples and cohorts. It should be noted that the negative sign here actually implies a positive ATE, because people born in the first three quarters $\mathcal{L}_{AK}^1$, $\mathcal{L}_{AK}^2$, and $\mathcal{L}_{AK}^3$ are associated with less years of schooling as compared

to those born in the fourth quarter. The CSL effect is strongest for those born in the first quarter and weakens with the second and third quarter born consecutively.

Table 5. Estimated average treatment effect (ATE) of CSL, $\beta_{y\mathcal{L}_s}$, using chain models (9) and (10).

	SY [a]								AK [b]			
	1930–1939				1940–1949				1930–1939		1940–1949	
	Full		School		Full		School		Full	School	Full	School
	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{AK}$		$\mathcal{L}_{AK}$	
$\mathcal{L}1$	0.025 **	0.011	0.022 *	0.011 *	0.054 **	0.016 *	0.048 **	0.016 **	−0.009 **	−0.007 **	−0.007 **	−0.007 **
	[12.5]	[2.56]	[6.25]	[3.40]	[88.4]	[6.26]	[125]	[20.8]	[99.2]	[75.2]	[112]	[95.4]
$\mathcal{L}2$	0.005	−0.003	0.021 *	−0.001	0.033 **	−0.004	0.045 **	−0.002	−0.006 **	−0.006 **	−0.004 **	−0.005 **
	[0.97]	[0.12]	[6.16]	[0.05]	[35.4]	[0.39]	[107]	[0.32]	[43.4]	[70.0]	[34.3]	[60.4]
$\mathcal{L}3$	0.004	0.012	0.011	0.023 **	0.027 **	0.002	0.028 **	0.014 **	−0.002 *	−0.002 *	−0.003 **	−0.002 **
	[0.44]	[3.86]	[1.56]	[15.9]	[20.3]	[0.12]	[37.8]	[11.4]	[4.62]	[5.68]	[15.1]	[6.00]

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. See Tables 3 and 4 for $\beta_{ys.Z}$ and β_{sL} estimates, respectively. Significance of $\beta_{y\mathcal{L}_s}$ based on χ^2 statistics estimated following Weesie (1999), reported in brackets. [a] Robust cluster adjusted standard errors. [b] Robust standard errors. ** Significant at the 1% level. * Significant at the 5% level.

Direct ATE estimates $\beta_{y\mathcal{L}}$ obtained via (7) exceed estimates obtained via chain approximation for the later born cohort; see Table 6. The effect is indicative of positive indirect CSL effects through control variables Z in later years. Further, chain approximations using SY indicators are much more varied across cohorts than across sub-samples, due to the varying estimates of β_{sL} in Table 3.

Table 6. Estimated ATE of the CSL $\beta_{y\mathcal{L}}$ via (7).

	SY [a]								AK [b]			
	1930–1939				1940–1949				1930–1939		1940–1949	
	Full		School		Full		School		Full	School	Full	School
	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{AK}$		$\mathcal{L}_{AK}$	
$\mathcal{L}1$	0.033	0.037 **	0.014	0.053 **	0.127 **	0.077 **	0.121 **	0.095 **	−0.014 **	−0.012 **	0.008 **	0.005
	(1.74)	(2.87)	(0.65)	(3.80)	(5.61)	(3.41)	(6.17)	(5.35)	(−4.23)	(−3.10)	(3.02)	(1.43)
$\mathcal{L}2$	−0.010	0.022	−0.0004	0.024	0.114 **	0.050 *	0.119 **	0.060 **	−0.009 **	−0.016 **	0.009 **	0.006
	(−0.62)	(1.48)	(−0.02)	(1.69)	(5.74)	(2.28)	(6.60)	(3.40)	(−2.87)	(−3.92)	(3.64)	(1.49)
$\mathcal{L}3$	0.014	0.071 **	0.006	0.106 **	0.110 **	0.095 **	0.105 **	0.125 **	0.0005	−0.005	0.010 **	0.016 **
	(0.77)	(4.85)	(0.31)	(6.49)	(5.75)	(3.88)	(5.90)	(6.03)	(0.17)	(−1.24)	(4.05)	(4.48)

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. t-statistics reported in parentheses. [a] Robust cluster adjusted standard errors. [b] Robust standard errors. ** Significant at the 1% level. * Significant at the 5% level.

Our finding of a moderate positive ATE of CSL on income (when using labour law indicators) is generally in line with findings reported in the literature; see Acemoglu and Angrist (2001); Lleras-Muney (2002); Oreopoulos (2006); and Goldin and Katz (2011).

5. What Have We Learnt?

Angrist and Pischke (2015, p. 227) discard the Acemoglu and Angrist (2001) study as 'a failed research design' and ascribe the failure to inappropriate CSL indicators, while maintaining the IV approach as appropriate. Conceptualising the IV approach as model choice and experimenting with the data sets used by AK and SY, our analysis exposes nescience about the causal model alternation nature of the IV approach to be the root cause of the failure instead.

Primarily, the model choice perspective enables us to transpose the IV-OLS choice into the selection between non-nested models with rival conditional variables. Since consistency is an asymptotic property, this selection can be assisted by CV methods from statistical learning. Our CV experiments show that the OLS-based models outperform the IV-based models in terms of generalisability and stability, regardless the choice of CSL indicators.

Careful examination of the causal implications of the CSL effects on ARTE by causal chain model representation helps us expose several logical flaws in the conceptualisation of IV-based models. First, it is incorrect to refer to β_{ys^L} as a consistent estimate of ARTE when $s^L s$. Second, the way in which s^L is generated entangles ARTE with the ATE by CSL in a non-unique manner, reaching deadlock in resolving the ambiguity over the causal interpretation of β_{ys^L}. Third, the argument for using IVs to treat measurement errors due to omission of correlated latent variables such as aptitude is unwarranted because ARTE is defined explicitly on education, not aptitude, which entails the specification of ARTA as the parameter of interest.

Experiments with models (9) and (10) show us that relatively robust $\beta_{ys.Z}$ estimates are attainable for ARTE, whereas this is not the case with various ATE estimates. The latter finding tells us that measurement error in CSL indicators is indeed a major concern, a result which confirms the common diagnosis of weak and/or inappropriate IVs in the literature. However, our results warn against the IV route as a dead-end in general when using IV to treat a latent variable problem since, in this case, measurement error in IVs is inevitable and also when prior knowledge suggests the need for explicit multivariate model specification with clear differentiation between moderator and mediator effects; see Arlot and Celisse (2010).

Finally, we find clear guidance and reassurance of our approach from the fundamental concepts and theories in statistical learning. In particular, model bias is identified as the primary source of model-based inferential bias. No theoretically postulated model should be taken as globally correct prior to empirical verification, and structural risk minimisation should be regarded as the key task of empirical studies. Applied research should thus be focused on agnostic probably approximately correct (PAC) learning. Once we fully recognise the untenability of the presumption of theoretically postulated models as globally correct, an implicit presumption underlying the IV-choice over OLS, the methodological defects of this estimator-centred research strategy transpire.

Appendix A

Complier Sub-Sample Experiment

Since most people remain in school beyond the required years, the great majority of the sample belongs to a sub-population for which the ATE of the CSL on schooling is expected to be 0. In other words, the CSL is potentially binding only for school leavers, but by and large not for those who have continued education beyond the compulsory years of schooling. Using $\mathcal{L}_{SY2}$ indicators, roughly 4.11% of the 1930–1939 born cohort complies[19] with the law. The share of compliers is even smaller for the later-born cohort with 2.31%. Using $\mathcal{L}_{SY2}$ indicators instead, the share of compliers is similarly small with 4.18 and 2.53% in the 1930s and 1940s birth cohorts, respectively (see Table A1). Our rough estimates of complier shares are slightly lower than in Bolzern and Huber (2017), who report a complier share of 6–12% for European countries based on comparison of mean potential outcomes using binary treatment and instrument variables.

---

[19] Compliers are overestimated here as the group includes some always takers that would have completed the years of schooling required by law, regardless of the law.

Table A1. Composition of CSL compliers, defectors, and always takers for $\mathcal{L}_{SY}$.

	Composition of CSL Compliers for SY $\mathcal{L}_{SY1}$					
	$\mathcal{L}^1_{SY1}=1$	$\mathcal{L}^2_{SY1}=1$	$\mathcal{L}^3_{SY1}=1$	**Total**	**Years of Schooling**	**Untreated**
	1930–1939					
Equal	1.37%	5.48%	4.10%	4.11%	$\mathcal{L}^1_{SY1}$: <7	3.75%
Less	2.09%	4.00%	10.19%	6.97%	$\mathcal{L}^2_{SY1}$: <8	5.88%
More	96.45%	90.57%	85.71%	88.50%	<9	12.28%
N	54,992	116,112	193,730	366,381	*N* (% of total)	1547 (0.42%)
	1940–1949					
Equal	0.45%	2.23%	2.90%	2.31%	$\mathcal{L}^1_{SY1}$: <7	3.20%
Less	0.69%	1.55%	5.40%	3.75%	$\mathcal{L}^2_{SY1}$: <8	5.28%
More	98.86%	96.22%	91.70%	91.71%	$\mathcal{L}^3_{SY1}$: <9	8.87%
N	86,252	112,481	337,979	548,870	*N* (% of total)	12,158 (2.22%)
	Composition of CSL Compliers for SY $\mathcal{L}_{SY2}$					
	$\mathcal{L}^1_{SY2}=1$	$\mathcal{L}^2_{SY2}=1$	$\mathcal{L}^3_{SY2}=1$	**Total**	**Years of Schooling**	**Untreated**
	1930–1939					
Equal	5.23%	4.11%	5.03%	4.18%	$\mathcal{L}^1_{SY2}$: <8	5.09%
Less	3.65%	10.67%	10.52%	7.44%	$\mathcal{L}^2_{SY2}$: < 9	10.79%
More	91.20%	85.22%	84.45%	88.38%	$\mathcal{L}^3_{SY2}$: < 10	14.73%
N	116,797	179,705	36,489	366,381	*N* (% of total)	33,390 (9.11%)
	1940–1949					
Equal	1.79%	3.00%	3.16%	2.53%	$\mathcal{L}^1_{SY2}$: <8	2.53%
Less	1.38%	5.64%	6.09%	4.30%	$\mathcal{L}^2_{SY2}$: <9	5.41%
More	96.83%	91.36%	90.74%	93.17%	$\mathcal{L}^3_{SY2}$: <10	8.37%
N	131,875	297,498	82,481	548,870	*N* (% of total)	37,016 (6.74%)

Notes: '*N*' is sample size, 'equal' is share of those with years of education equal to school law, and 'less' years of education and 'more' years of education, respectively, for those treated by the respective law. For the untreated group, share of those with less than 7, 8, and 9 years of education among untreated is given. 'Total' compares 'Untreated' against total of the sample in the equal to the law, less than the law, and more than the law of schooling categories.

Following from the above, the difference in the IV estimates for the ARTE when using different CSL indicators is commonly explained to be the result of localised treatment, i.e., treatment which is confined to specific complier groups; see Angrist et al. (1996) and Angrist and Imbens (1995) more generally. We utilise these insights and examine the localness using sub-sample data. We replicate SY Tables 1 and A2 and AK Tables V and VI using sub-samples to separate 'always takers' from 'compliers', drawing on the above considerations underlying Table A1.[20] Specifically, those who receive 12 or less years of schooling are allocated to the *School* sub-sample and those with more years of education are allocated to the *Higher* sub-sample. Further, the tails of the two sub-samples, *Higher* and *School*, are cut to investigate whether dissimilarities between the sub-samples arise due to outliers (cf. Figure A1). The aim of this experiment is to examine the empirical consistency of the two models under the localised treatment condition.

The IV model estimates are close to the OLS counterparts for most of the *School* sub-sample, while β^L is insignificant for the *Higher* sub-sample, confirming our conjecture of 0 ATE by CSL treatment for the latter sub-sample (Table A2).[21] The only IV-based model specification that yields $\beta^L \neq \beta$ with instruments that are not refuted across the two cohorts is based on $\mathcal{L}_{SY1}$ indicators for the *School* sub-sample. Interestingly, the OLS estimates, while significant and positive throughout all sub-samples

---

[20] The separation remains imperfect, as some always takers will be contained in the complier group.
[21] The only exception is the later-born cohort with AK's model specification, where all return to education estimators are insignificant or diagnostics reveal problems with the model design.

and cohorts, vary across sub-samples, with the ARTE decreasing for those attaining 13 to 15 years of education.[22]
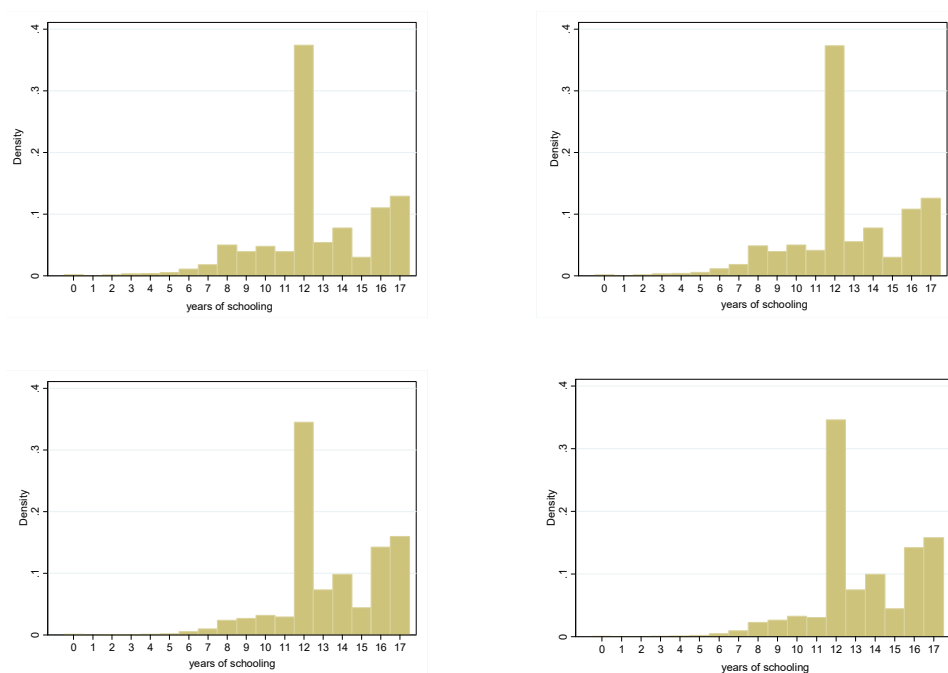

Figure A1. Years for schooling density. Notes: AK education variable is capped at 17 years of schooling for comparability between the AK and SY data sets.

Table A2. Estimation of β via sub-sampling on educational attainment.

	1930–1939 Born				1940–1949 Born			
Years of	**School**		**Higher**		**School**		**Higher**	
Schooling	**≤12**	**7–12**	**13–15**	**≥13**	**≤12**	**7–12**	**13–15**	**≥13**
	SY							
β (SY)	0.0613 **	0.0587 **	0.0546 **	0.0935 **	0.0612 **	0.0596 **	0.0273 **	0.0641 **
[95% CI] [a]	[0.0602–0.0625]	[0.0573–0.0601]	[0.0476–0.0615]	[0.0912–0.0957]	[0.0599–0.0624]	[0.0581–0.0611]	[0.0223–0.0322]	[0.0625–0.0657]
β^L (with $\mathcal{L}_{SY1}$)	0.0942 **	0.0860 *	−0.2577	−0.0218	0.1008 **	0.2072 **	0.4449	0.2292
[95% CI] [a]	[0.0623–0.1261]	[0.0094–0.1626]	[−0.809–0.2937]	[−0.585–0.541]	[0.0710–0.1305]	[0.0558–0.3586]	[−0.101–0.991]	[−0.124–0.583]
Sargan	0.66 (0.7198)	0.17 (0.9173)	2.43 (0.2963)	5.44 (0.0660)	2.07 (0.3549)	3.59 (0.1661)	8.58 (0.0137)	22.7 (0.0000)
Hausman	4.07 (0.0437)	0.49 (0.4859)	1.43 (0.2325)	0.17 (0.6805)	7.02 (0.0081)	4.63 (0.0315)	2.93 (0.0867)	0.98 (0.3230)
β^L (with $\mathcal{L}_{SY2}$)	0.1596 **	0.1411 *	−0.5833	0.0831	0.0836 **	0.0671	0.3829	0.0168
[95% CI] [a]	[0.1182–0.2010]	[0.0193–0.2630]	[−1.76–0.591]	[−0.341–0.5073]	[0.0499–0.1172]	[−0.030–0.164]	[−0.203–0.969]	[−0.267–0.3004]
Sargan	0.21 (0.8990)	11.7 (0.0028)	3.06 (0.2171)	0.56 (0.7563)	7.27 (0.0264)	4.23 (0.1209)	2.33 (0.3123)	9.23 (0.0099)
Hausman	24.6 (0.0000)	1.86 (0.1732)	1.79 (0.1807)	0.002 (0.962)	1.72 (0.1896)	0.02 (0.8799)	1.75 (0.1862)	0.11 (0.7419)
β (AK)	0.0565 **	0.0583 **	0.0442 **	0.0591 **	0.0701 **	0.0734 **	0.0153 **	0.0405 **
[95% CI]	[0.0551–0.0578]	[0.0564–0.0601]	[0.0371–0.0515]	[0.0575–0.0607]	[0.0686–0.0716]	[0.0715–0.0753]	[0.0105–0.0201]	[0.0393–0.0416]
β^L (with $\mathcal{L}_{AK}$)	0.0864 **	0.1107 **	0.0295	−0.0446	0.0067	−0.0040	0.0606	0.2612 **
[95% CI]	[0.0357–0.1372]	[0.0288–0.1926]	[−0.309–0.368]	[−0.147–0.058]	[−0.051–0.0641]	[−0.078–0.0696]	[−0.268–0.3888]	[0.1580–0.3644]
Sargan	24.4 (0.7088)	24.2 (0.7217)	33.4 (0.2633)	30.1 (0.4097)	60.8 (0.0005)	59.6 (0.0007)	38.0 (0.1230)	33.5 (0.2587)
Hausman	1.35 (0.2447)	1.60 (0.2054)	0.01 (0.9320)	4.44 (0.0352)	4.84 (0.0279)	4.37 (0.0367)	0.07 (0.7863)	27.5 (0.0000)

Notes: SY data follows model specification (1) Tables 1 and A2; AK data follows Tables V and VI columns (5) and (6) model specifications. The discrepancy between OLS estimates obtained with SY and AK data sets is due to differences in the construction of the education variable; see Table 3 and discussion. *p*-values for Sargan and Hausman test in parentheses. [a] Standard errors cluster adjusted. ** Significant at the 1% level. * Significant at the 5% level.

---

[22] This effect is more pronounced for the later-born cohort, potentially due to educational inflation. These data patterns are undetectable by the CSL-based IV method, since instruments narrowly target school goers but not those attaining higher education and large standard errors hide significant difference across cohorts.

Appendix B

Testing for a CSL-Induced Shift in the Schooling Effect: $\beta_{ys_L.Z} \neq \beta_{y\widetilde{s}_L.Z}$

　　　If the introduction of the CSL has altered the ability composition of workers, the post-treatment schooling effect, $\beta_{ys_L.Z}$, might differ from the schooling effect before treatment, $\beta_{y\widetilde{s}_L.Z}$, namely $\beta_{ys_L.Z} \neq \beta_{y\widetilde{s}_L.Z}$. To investigate if such a parameter shift has occurred, break-point Chow tests are conducted. The test requires careful sub-sample division conditional on the *L*-treatment and also different choices of CSL indicators.

　　　For the AK law indicators, the treated group is defined as those born in the first and second quarters of the year, while the untreated group is defined as those born in the remaining quarters. Since the CSL is binding for only a minority of the treated group—as evident from the negligible share of compliers (Table A1, Appendix A)[23]—the ability composition is unlikely to render a significant parameter shift using the full sample. As before, sub-sample division is used to minimises defectors and always takers in the treated group of the *School* sub-sample. A parametric shift should hence be discernible for the *School* but not the *Higher* sub-sample. As can be seen from Table A3, despite the careful sub-sample division, the null hypothesis of no break point cannot be rejected at the 5% level in all experimental settings.

　　　For the SY indicators, the sub-sample division is slightly more complicated. $\mathcal{L}_{SY}$ indicators capture the minimum years of schooling and school attendance required by the respective state's labour and education law, but only few states had no law in place over the sample periods, which results in a rather small sub-sample for the untreated. Further, the nature of the CSL indicators enables us to minimise defectors and always takers even further in the treated group of the *School* sub-sample. The treated group is re-defined as those among the *School* sub-sample who received treatment under a particular law and leave school right after the minimum years of schooling required by the law are completed. The sub-sample is denominated *School Binding.* Although the treated group does not fully capture those compliant to the law, the group fully encompasses those compliant while minimising the number of those defecting in the group under the given information.

Table A3. Test for $\mathcal{L}_{AK}$ -induced parametric shift.

	1930–1939			1940–1949		
	Full	**School (7–12)**	**Higher (13–15)**	**Full**	**School (7–12)**	**Higher (13–15)**
Chow Test	0.65	0.72	0.19	0.76	0.25	0.18
(Treated-Full) [a]	(0.4184)	(0.3949)	(0.6652)	(0.3842)	(0.6137)	(0.6673)
Chow Test	0.56	0.63	0.19	1.05	0.17	0.18
(Untreated-Full) [b]	(0.4549)	(0.4286)	(0.6617)	(0.3053)	(0.6769)	(0.6673)

Notes: 1980 census, males with positive weekly earnings. *p*-values in parentheses. [a] The treated sub-sample is defined as those born in the first and second quarters. [b] Untreated sub-sample is defined as those born in the third and fourth quarters.

　　　As can be seen from Table A4, using $\mathcal{L}_{SY1}$ indicators, the break-point Chow tests provide no evidence for a treatment induced parametric shift in the ARTE parameter at the 1% significance level, and some evidence at 5% significance level for the *School Binding* sub-sample. Using $\mathcal{L}_{SY2}$ as indicator, there is some evidence for a parametric effect for the later born cohort. The effect is absent from the *Higher* sub-sample as expected, but detectable at the 5% level for the *Full* sample and *School* sub-sample. However, given measurement errors identified for the $\mathcal{L}_{SY2}$ indicators in Table 4, the evidence is too weak to conclude on a parametric shift.

---

[23]　Although Table A1 (Appendix A) is based on SY data, the patterns are likely to be identical for the AK data.

Table A4. Test for $\mathcal{L}_{SY}$ -induced parametric shift in $\beta_{ys_L.Z}$

	1930–1939							
	Full		School (7–12 y)		Higher (13–15 y)		School Binding [c]	
	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$
Chow Test	0.01	1.80	0.01	0.06	0.46	0.09	1.26	2.59
(Treated-Full) [a]	(0.9172)	(0.1797)	(0.9236)	(0.8056)	(0.4990)	(0.7693)	(0.2607)	(0.1076)
Chow Test	0.02	0.37	0.03	0.10	0.38	0.20	4.45 *	0.19
(Untreated-Full) [b]	(0.8848)	(0.5417)	(0.8574)	(0.7556)	(0.5395)	(0.6569)	(0.0350)	(0.6596)
	1940–1949							
	Full		School (7–12 y)		Higher (13–15 y)		School Binding [c]	
	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$	$\mathcal{L}_{SY1}$	$\mathcal{L}_{SY2}$
Chow Test	2.24	2.28	1.49	1.96	0.02	0.00	0.00	14.05 **
(Treated-Full) [a]	(0.1342)	(0.1312)	(0.2220)	(0.1617)	(0.8847)	(0.9896)	(0.9580)	(0.0002)
Chow Test	5.04	4.23 *	3.25	6.06 *	0.05	0.02	0.26	2.41
(Untreated-Full) [b]	(0.0248)	(0.0396)	(0.0713)	(0.0138)	(0.8195)	(0.8966)	(0.6115)	(0.1206)

Notes: 1980 census, white males with positive weekly earnings. *p*-values in parentheses. [a] The treated sub-sample is defined as those born in a state with some school law in place, that is, minimum years of schooling unequal to 0. [b] Untreated sub-sample is defined as those born in a state with no school law in place, that is, minimum years of schooling equal to 0. [c] Definition of treated and untreated change for this experiment. The treated sub-sample is defined as those who drop out after the minimum years of schooling and the untreated sub-sample comprises of the remaining observations. ** Significant at the 1% level. * Significant at the 5% level.

Overall, the evidence for the presence of an L-treatment-induced parametric shift is weak and rattled with deficiencies in the CSL indicators. However, if the substantive interest is with a data-consistent ARTE effect, data patterns originating from sheepskin effects and educational inflation are found to be of far greater concern than an L-treatment-induced parametric shifts (see Table A2 Appendix A and Table 2).

References

Abadie, Alberto, and Matias D. Cattaneo. 2018. Econometric Methods for Program Evaluation. *Annual Review of Economics* 10: 465–503. [CrossRef]

Acemoglu, Daron, and Joshua D. Angrist. 2001. How Large Are Human Externalities? Evidence from Compulsory Schooling Laws. *NBER Macroeconomics Annual* 15: 9–74. [CrossRef]

Angrist, Joshua D. 1995. The Economic Returns to Schooling in the West Bank and Gaza Strip. *The American Economic Review* 85: 1065–87.

Angrist, Joshua D., and Guido W. Imbens. 1995. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association* 90: 431–42. [CrossRef]

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91: 444–55. [CrossRef]

Angrist, Joshua D., and Alan B. Krueger. 1991. Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics* 106: 979–1014. [CrossRef]

Angrist, Joshua D., and Alan B. Krueger. 1992. The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of the American Statistical Association* 87: 328–36. [CrossRef]

Angrist, Joshua D., and Joern-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect, Princeton.* Princeton and Woodstock: Princeton University Press.

Angrist, Joshua D., and Joern-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion, Princeton.* Princeton and Woodstock: Princeton University Press.

Arlot, Sylvain, and Alain Celisse. 2010. A Survey of Cross-Validation Procedures for Model Selection. *Statistical Surveys* 4: 40–79. [CrossRef]

Athey, Susan, and Guido W. Imbens. 2015. *Machine Learning Methods for Estimating Heterogeneous Causal Effects, Mimeo.* Stanford: Stanford University.

Bolzern, Benjamin, and Martin Huber. 2017. Testing the Validity of the Compulsory Schooling Law Instrument. *Economics Letters* 159: 23–27. [CrossRef]

Bound, John, and David A. Jaeger. 2000. Do Compulsory School Attendance Laws Alone Explain the Association between Quarter of Birth and Earnings? *Worker Well-Being* 19: 83–108.

Card, David. 2001. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica* 69: 1127–60. [CrossRef]

Card, David, and Alan B. Krueger. 1992a. Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy* 100: 1–40. [CrossRef]

Card, David, and Alan B. Krueger. 1992b. School Quality and Black-White Relative Earnings: A Direct Assessment. *Quarterly Journal of Economics* 107: 151–200. [CrossRef]

Carneiro, Pedro, and James J. Heckman. 2002. The Evidence on Credit Constraints in Post-Secondary Schooling. *The Economic Journal* 112: 989–1018. [CrossRef]

Clark, Damon, and Paco Martorell. 2014. The Signaling Value of a High School Diploma. *Journal of Political Economy* 122: 282–318. [CrossRef]

Cox, David R., and Nanny Wermuth. 1996. *Multivariate Dependencies–Model, Analysis and Interpretation*. London: Chapman & Hall.

Cox, David R., and Nanny Wermuth. 2004. Causality: A Statistical View. *International Statistical Review* 72: 285–305. [CrossRef]

Deaton, Angus. 2009. *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*. NBER Working Paper Series; No. 14690. Cambridge: NBER.

Deaton, Angus, and Nancy Cartwright. 2018. Understanding and Misunderstanding Randomized Controlled Trials. *Social Science & Medicine* 210: 2–21.

Durbin, James. 1954. Errors in Variables. *Review of the International Statistical Institute* 22: 23–32. [CrossRef]

Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. *Econometrica* 51: 277–304. [CrossRef]

Entner, Doris, Patrik O. Hoyer, and Peter Spirtes. 2012. Statistical Test for Consistent Estimation of Causal Effects in Linear Non-Gaussian Models. Paper presented at the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands, Spain, April 21–23.

Goldin, Claudia. 1998. America's graduation from high school: The evolution and spread of secondary schooling in the twentieth century. *Journal of Economic History* 58: 345–74. [CrossRef]

Goldin, Claudia, and Lawrence F. Katz. 2000. Education and Income in the Early 20th Century: Evidence from the Prairies. *The Journal of Economic History* 60: 782–818. [CrossRef]

Goldin, Claudia, and Lawrence F. Katz. 2007. *Long-Run Changes in the U.S. Wage Structure: Narrowing, Widening, Polarizing*. NBER Working Paper Series; No. 13568. Cambridge: NBER.

Goldin, Claudia, and Lawrence F. Katz. 2011. Mass Secondary Schooling and the State: The Role of State Compulsion in the High School Movement. In *Understanding Long Run Economic Growth*. Edited by Dora Costa and Naomi Lamoreaux. Chicago: University of Chicago Press, pp. 275–310.

Harmon, Colm, Hessel Oosterbeek, and Ian Walker. 2003. The Returns to Education: Microeconomics. *Journal of Economic Surveys* 17: 115–55. [CrossRef]

Heckman, James J., and Sergio Urzua. 2009. *Comparing IV with Structural Models: What Simple IV Can and Cannot Identify*. NBER Working Paper Series; No. 14706. Cambridge: NBER.

Hoogerheide, Lennart, and Herman K. van Dijk. 2006. *A Reconsideration of the Angrist-Krueger Analysis on Returns to Education*. Econometric Institute Report EI 2006-15. Rotterdam: Econometric Institute.

Imbens, Guido W. 2010. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48: 399–423. [CrossRef]

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. Identification and Inference with Many Invalid Instruments. *Journal of Business & Economic Statistics* 33: 474–84.

Lleras-Muney, Adriana. 2002. Were Compulsory Attendance and Child Labor Laws Effective? An Analysis from 1915 to 1939. *The Journal of Law and Economics* 45: 401–35. [CrossRef]

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Roland C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2013. Long-Term Neighborhood Effect on Low-Income Families: Evidence from Moving to Opportunity. *American Economic Review* 103: 226–31. [CrossRef]

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Roland C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2012. Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science* 337: 1505–10. [CrossRef]

Mukherjee, Sayan, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. 2006. Learning Theory: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization. *Advances in Computational Mathematics* 25: 161–93. [CrossRef]

Murphy, Kevin M., and Finis Welch. 1992. The Structure of Wages. *The Quarterly Journal of Economics* 107: 285–326. [CrossRef]

Oreopoulos, Philip. 2006. Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter. *American Economic Review* 96: 152–75. [CrossRef]

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Qin, Duo. 2015. Resurgence of the Endogeneity-Backed Instrumental Variable Methods. *Economics: The Open-Access, Open-Assessment E-Journal* 9: 1–35. [CrossRef]

Qin, Duo. 2018. Let's Take the Bias Out of Econometrics. *Journal of Economic Methodology* 26: 81–98. [CrossRef]

Qin, Duo, Sophie van Huellen, Raghda Elshafie, Yimeng Liu, and Thanos Moraitis. 2019. *A Principled Approach to Assessing Missing-Wage Induced Selection Bias*. SOAS Department of Economics Working Paper Series; No. 216. London: SOAS.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2010. Learnability, Stability and Uniform Consistency. *Journal of Machine Learning Research* 11: 2635–70.

Stephens, Melvin, Jr., and Dou-Yan Yang. 2014. Compulsory Education and the Benefits of Schooling. *American Economic Review* 104: 1777–92. [CrossRef]

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: 518–29.

Trostel, Philip. A. 2005. Nonlinearity in the Return to Education. *Journal of Applied Economics* 8: 191–202. [CrossRef]

Weesie, Jeroen. 1999. Seemingly Unrelated Estimation and The Cluster-adjusted Sandwich Estimator. *Stata Technical Bulletin* 52: 34–47, Reprinted in *Stata Technical Bulletin Reprints* 9: 231–48.

Wermuth, Nanny, and David R. Cox. 2011. Graphical Markov Models: Overview. In *International Encyclopedia of Social and Behavioral Sciences*, 2nd ed. Edited by Wright James. New York: Elesevier, pp. 341–50.

Young, Alwyn. 2017. *Consistency without Inference: Instrumental Variables in Practical Application, Mimeo*. London: London School of Economics.

Zhang, Kun, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. 2017. Causal Discovery in the Presence of Measurement Error: Identifiability Conditions. Paper presented at the UAI 2017 Workshop on Causality, Sydney, Australia, August 11–15.

Zhang, Yongli, and Yuhong Yang. 2015. Cross-Validation for Selecting a Model Selection Procedure. *Journal of Econometrics* 187: 95–112. [CrossRef]