Compulsory Schooling and the Returns to Education: A Re-examination

by

Sophie van Huellen and Duo Qin

(November, 2018)
The **SOAS Department of Economics Working Paper Series** is published electronically by The School of Oriental and African Studies-University of London.

©Copyright is held by the author or authors of each working paper. SOAS DoEc Working Papers cannot be republished, reprinted or reproduced in any format without the permission of the paper’s author or authors.

This and other papers can be downloaded without charge from:

**SOAS Department of Economics Working Paper Series** at
[http://www.soas.ac.uk/economics/research/workingpapers/](http://www.soas.ac.uk/economics/research/workingpapers/)

Research Papers in Economics (RePEc) electronic library at
[http://econpapers.repec.org/paper/](http://econpapers.repec.org/paper/)

Design and layout: O.G. Dávila
Compulsory Schooling and the Returns to Education: A Re-examination

Sophie van Huellen* and Duo Qin**

Abstract

This paper re-examines the instrumental variable approach to estimating the effect of compulsory school law on education in the US pioneered by Angrist and Krueger (1991). We show that the approach not only yields empirically inconsistent estimates but is conceptually confused. The confusion arises from a rejection of the key causal variable as a valid conditional variable. By route of a causally explicit model design assisted by machine learning techniques, we identify the circumstances under which the wrongly rejected variable yields valid inference values. Our investigation demonstrates the importance of data-guided model selection over the choice of consistent estimators.

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

JEL classification: C26, C52, I21, I26, J24

* Department of Economics, SOAS, University of London. Email: sv8@soas.ac.uk  ** Department of Economics, SOAS, University of London. Email: dq1@soas.ac.uk

This paper has been updated from an earlier version published as working paper in September 2016.
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 are pioneered by Angrist and Krueger (1991). They use CSL indicators as instrumental variables (IVs) to ‘randomize’ latent ability across educational attainment groups in order to correct for the assumed bias in the ordinary least square (OLS) estimator; see also and Acemoglu and Angrist (2001). The influence of this seminal paper reaches far beyond the estimation of CSL effects. The empirical strategy is now common practice in research on the average return to education (ARTE) as well as program evaluation modelling, e.g. Harmon et al. (2003), Ludwig et al. (2012; 2013), and the two papers have since entered the standard economics curriculum as evident from their appearance in two popular textbooks by Angrist and Pischke (2009; 2015).

Despite the far-reaching influence of these studies, causal interpretation of the estimated CSL effect is confused. This is exemplified by two interlinked developments in the literature: a) a shift in the interpretation of the CSL instrumentalized returns to schooling coefficient despite identical modelling choice, and b) empirical results which vary significantly with the choice of CSL indicators. Angrist and Krueger (1991), who approximate CSL with quarter of birth dummies, interpret their results as unbiased estimates of the ARTE, and find that the IV estimates are not statistically different from estimates obtained via OLS.\(^1\) Acemoglu and Angrist (2001) construct alternative CSL indicators based on labor law that produce IV estimates which, although significantly different from OLS estimates, are insignificant or negative.\(^2\) Acemoglu and Angrist (2001) no longer interpret the IV estimates as unbiased returns to schooling, but as the causal effect of CSL on earnings via schooling. These CSL indicators are further refined by Stephens Jr. and Yang (2014). Their results verify the findings by Acemoglu and Angrist (2001) and Stephens Jr. and Yang (2014: 1789) conclude that there is ‘no evidence of benefits to additional schooling’ due to CSL.

This paper draws upon concepts and techniques of causal modelling and model selection in statistical machine learning to probe into the confusion over the CSL instrumentalized ARTE estimator. By careful re-examination of two data sets from a causal modelling angle, one from

\(^1\) E.g. column (5) versus (6) in Table 4, (7) versus (8) in Table 5, (1) versus (2) and (5) versus (6) in Table 6 in Angrist and Krueger (1991). More evidence in Hoogerheide and van Dijk (2006, Table 5) and in Harmon et al. (2003, Section 5).
\(^2\) See Angrist and Pischke (2015, Table 6.3) for a summary.
Angrist and Krueger (1991) (AK hereafter) and the other from Stephens Jr. and Yang (2014) (SY hereafter), \(^3\) we expose the methodological obfuscations underlying the confusion.

We start from the root of the confusion – causal model specification. Close inspection of the causal links between the explanatory variables of interest with the help of causal graphs in Section 2 reveals two key insights. First, the use of IV estimators amounts to making, albeit implicitly, an *a priori* assertion of the education variable being an invalid conditional variable, thus undermining its causal status. We reason that the ensuing replacement of the education variable by an instrumented alternative amounts to a change in model specification. Conceptualizing the IV approach as model choice, the implicit assertion underlying the use of IV estimators can be explicitly specified into a testable hypothesis. Second, two types of outcome effects are carefully distinguished via causal chain model estimation: ARTE with a possible moderation effect of CSL and the average treatment effect (ATE) of the CSL via schooling.

These two insights are empirically evaluated in Sections 3 and 4. The validity of the implicit assertion underlying the IV approach is overwhelmingly rejected by k-fold cross validation (CV), an essential tool for model comparison in machine learning, in Section 3. Further examination into the circumstances where the education variable constitutes a valid conditional variable yields positive results in Section 4. Specifically, the decomposition of the two outcome effects reveals a virtually invariant and empirically consistent ARTE estimate of 0.06, and smaller ATE of the CSL estimates between 1-5% if using labor law indicators and 0.2-0.9% if using quarter of birth indicators. Unfortunately, CSL outcome and moderation effects are undermined by measurement errors in the indicators provided in AK and SY.

The empirical results in Section 4 open the door for a causally explicit model design through statistical data learning. The methodological implications are extended in Section 5. Angrist and Pischke (2015: 227) discard the Acemoglu and Angrist (2001) study as ‘a failed research design’ and ascribe the failure to the choice of inappropriate CSL indicators that fail to meet identifying assumptions; see also Stock et al. (2002), Kolesár et al. (2015). In contrast, our analysis reveals the root of the failure to be the choice of the IV approach, not the choice of instruments. We argue that the IV approach is a detour into deadlock, as it effectively denies direct translation of causal postulates of interest into data-consistent conditional relationships; an argument that applies to the use of the IV approach as a panacea for endogeneity bias in

\(^3\) The two datasets 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).
general; see Qin (2019). Our re-examination of the CSL case clearly shows the importance of empirical model design via data learning over estimator choice for untested theoretical models.

2. Model Specification of Schooling Effects Under CSL Treatment: An Anatomy

To disentangle the conflicting causal interpretations presented in AK and SY, let us start from their basic model design. Briefly, denote education by \( s \) and outcome by \( y \), the ARTE is represented by \( \beta \) in (1):

\[
y = \alpha + \beta s + \eta.
\]

AK and SY assume \( \text{cov}(s\eta) \neq 0 \), based on the argument that \( \eta \) contains omitted variables which are not directly observable but collinear with \( s \), such as innate ability. The assumption determines OLS to be an inconsistent estimator for (1) which should be replaced by consistent IV estimators. AK and SY use the CSL as a key instrument. The resulting IV estimator amounts to appending (1) by:

\[
s = \pi L + I_j' \gamma_j + \varepsilon,
\]

where \( L \) represents the CSL and \( I_j \) other IVs. Equation (2) can be seen as the first stage of the two-stage least square (2SLS) estimation. Combining it with (1) and taking into consideration the shift in the interpretation of \( \beta \) from the OVB-free ARTE to the CSL effect via schooling, we see that the IV approach actually implies a rejection of (1) in favor of:

\[
y = \alpha + \beta^L s^L + \eta^L,
\]

where \( s^L \) denotes the fitted \( s \) from (2) and superscript \( L \) represents the CSL. The expected \( \beta^L \neq \beta \) entails \( s^L \neq s \), which leads to the shifted interpretation. Hence, the essence of the IV approach is to assert \( s^L \), instead of \( s \), as the valid conditional variable – an implicit change in model specification. In other words, the choice between OLS and IV estimators is essentially a choice of a valid conditional variable of interest among \( s \) and \( s^L \), and consequently a valid conditional model among (1) and (3); see Qin (2015; 2019). Motivated by this insight, we proceed with an anatomy of model specification that goes beyond the narrow estimation perspective.

\*A detailed exposition of the methodological flaws of this assumption is given in Qin (2019).
In order to understand what has caused the assertion of $s$ being an invalid conditional variable and the shift in interpretation of $\beta^L$, we further scrutinize schooling effects under the CSL treatment. Since the CSL effect on income via schooling is a sequential event, an intuitive way to represent the event is a directed acyclic graph (DAG), a popular tool for assisting causal modelling; see Cox and Wermuth (1996), Wermuth and Cox (2011) and Pearl (2009).

![DAGs of Returns to Schooling Under the CSL Treatment](image)

**Figure 1. DAGs of Returns to Schooling Under the CSL Treatment**

*Notes: y denotes earnings, s, schooling, L, the CSL and L its observable indicator. A node inside a square indicates a latent variable; a solid node denotes a dummy/binary variable.*

The DAG in the left panel of Figure 1 shows us that, when the effect of $L$ forms the focal causal interest, $s$ takes the role of an intermediate variable, or a mediator, but when $s$ is the causal variable of interest, $L$ takes the role of a moderator exclusive for $s$. Whatever the causal variable of interest, $f(y, s, L)$, the joint density of the three variables can be recursively factorized and reduced as below:

$$f(y, s, L) = f(y|s, L)f(s|L)f(L) = f(y|s, L)f(s|L).$$

since $f(L) = 1$ when retrospective cross-section data samples are used. From Pearl (2009), it can be seen that the CSL treatment implies a rule of intervention, namely $y \perp L|s$, so that we can further factorize the conditional density in (4) as:

$$f(y, s|L) = f(y|s)f(s|L).$$

The sequential nature of the ATE of $L$ on $y$ via $s$ is expressed by the conditional expectation of (5), $E(y, s|L) = E(y|s)E(s|L)$. In a linear model setting, the ATE, denoted by $\beta_{yL}$, can be derived from a chain of two simple regressions corresponding to $E(y|s)$ and $E(s|L)$ respectively; see Cox and Wermuth (2004).  

---

5 Refers to Rule 2 on the external intervention in (Pearl 2009, Section 3.4). This is equivalent to the exclusion restriction in the IV literature.

6 See Cox and Wermuth (2004) for a general discussion on the parametric chain representation. We adopt their method of subscript-based parametric notation to highlight the consequence of different causal chain specifications for the parameters of regressors.
Model (6) tells us that $\beta_{yL} \neq \beta_{ys}$ holds 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 $\beta_{yL} \ll \beta_{ys}$.

If $\beta_{yL}$ forms the only parameter of substantive interest, the chain route appears a long way round to estimate it, because the parameter can be estimated directly from:

\begin{equation}
(7) \quad y = a_y + \beta_{yL}L + \epsilon_y.
\end{equation}

However, $L$ is latent in the cross-section samples used by AK and SY and is approximated by various observable indicators, $L$, i.e.:

\begin{equation}
(7') \quad y = a_y + b_{yL}L + \epsilon_y.
\end{equation}

This direct route may result in $b_{yL} \neq \beta_{ys}\beta_{sl}$. The defectiveness of CSL indicators due to entanglement with regional factors and other controls leading to indirect effects has already been identified by SY. In other words, $L$ may fail the rule of intervention such that $\beta_{yL} \neq 0$ from the following regression:

\begin{equation}
(8) \quad y = a_y + \beta_{yL}L + \epsilon_y.
\end{equation}

This situation is illustrated in the modified DAG in the right panel of Figure 1. Two consequences follow. First, the chain route of (6) is more reliable than (7') for estimating the ATE of the CSL via schooling. Second, a test of $\beta_{yL} = 0$ using (8) can be exploited as an additional criterion for CSL indicator selection purpose; see Zhang et al. (2017) for implication of measurement error for the estimation of causal chains.

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 $\text{cov}(sZ) \neq 0$, some of the control variables are likely to correlate with CSL indicators, 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 between $Z$ and CSL indicators complicates the estimation of the CSL effect. Extend (6) by $Z$: 

(6) \[ y = \alpha_y + \beta_{ys}s + \epsilon_y \]

\[ s = \alpha_s + \beta_{sl}L + \epsilon_s \Rightarrow \beta_{yL} = \beta_{ys}\beta_{sl}. \]
\[ y = \alpha_y + \beta_{ys} s + Z' \beta_{yz} s + \varepsilon_y \]
\[ s = \alpha_s + \beta_{zl} L + \varepsilon_s \]
\[ Z = \alpha_Z + L \beta_{zl} + \varepsilon_Z \]

The corresponding chain representation of parameters becomes:

\[ (10) \beta_{yl} = \beta_{ys} \beta_{zl} + \beta_{yz} \beta_{zl} = \beta_{ys} + \beta_{yl} \]

Now, only the first component, \( \beta_{yl} \), in (10) corresponds to the ATE of the CSL via schooling. The simple regression model of (7') is no longer fit for the purpose.

Model (9) differs from model (3) in two essential aspects. First, \( s \) is taken as a valid conditional variable in the first equation of (9). Second, the lower equations are part of the chain representation rather than an instrumental step. The two income effects can thus be estimated separately.

The separation of the two effects enables us to disentangle the potential problems arising in their estimation. The OVB-based argument underlying the IV-treatment applies to \( \beta_{ys} \) in (9) because of the inclusion of \( Z \). \( \beta_{yl} \) or \( \beta_{yls} \) defines OVB or part of it if viewed from \( \beta_{yl} \) of (7) when estimating the CSL effect. Hence, the question is, whether \( \beta_{ys} \) is consistent when controlling for a set of covariates \( Z \).

Further, concerns over measurement-error are far more relevant to CSL indicators than to \( s \) in the present context.\(^7\) However, a concern over selection bias remains pertinent to the estimation of the schooling effect. Specifically, the 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: Chapter 4),

---

\(^7\)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).
so much so that the post-treatment schooling effect becomes significantly different from the pre-treatment one.

The IV remedy of this selection-bias induced effect is to maintain $\beta^t \neq \beta$ from the perspective of retrospective data. As pointed out earlier, the IV approach amounts to rejecting the causal validity of $s$ and substituting it with $s^L$. Figure 3 illustrates the situation in two DAGs.

![Figure 3. DAGs of Model (1) (Left) versus Model (3) (Right)](image)

*Notes*: 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), from a chain model specification with respect to $L$.

Comparison of Figure 3 with Figure 1 reveals a key difference between models (3) and (8) as different alternatives to (1). While (8) extends (1) via a causal chain model specification as illustrated in the left panel of Figure 1, the IV route by (2) confines $L$ to an instrumental role in producing a new variable $s^L$, whereby blocking the route to a chain model extension.\(^8\)

Consequently, the IV approach is unsuitable for tackling the question whether the CSL treatment has indeed resulted in a compositional shift to such an extent that it has caused a parametric shift in the ARTE, since $\beta$ of (1) is already rejected *a priori* as an inconsistent parameter. Under the chain model specification, the only feasible way to tackle this question is 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: $\beta_{yLS} \neq \beta_{yS^LZ}$, where $s_L$ denotes schooling of the $L$-treated part, and $S_L$ the treatment unaffected part. It should be noted that even if the inequality is supported by data, the evidence alone is insufficient for rejecting $s$ as a valid conditional variable for $y$, e.g. see Engle et al. (1983). Conversely, evidence against such a parametric shift does not imply $\beta_{yLZ} = 0$ unless $\beta_{SL} = 0$ can be verified.

### 3. Where Does the Research Design Fail?

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. This re-specification is based on the premise of

---

\(^8\) 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).
inconsistency of the OLS model specification relative to its IV counterpart. By exposing the IV approach as a model re-specification, the choice of the key conditional variable and therefore the choice between the IV and OLS model becomes a question of model selection which is testable. According to statistical machine learning theory, model selection is based on empirical risk minimization 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 generalizability (or predictivity), stability and consistency, whereby Mukherjee et al. (2006) show that stability is equivalent to empirical consistency. One common tool for model selection is CV, which provides a formal assessment of predictivity and consistency by splitting the sample into a training and testing component, e.g. see Arlot and Celisse (2010). The tool is not new to the impact evaluation literature, e.g. Athey and Imbens (2015).

In the following, we first replicate and then, by use of CV, reassess the results presented by AK and SY against the above listed criteria. Since the CSL is latent, it is approximated by observable indicators, $L$. Quarterly birth dummies are chosen by AK ($L_{AK}$).\(^9\) SY, with reference to Acemoglu and Angrist (2001), propose two alternative indicators based on state school and labor law. These indicators capture required years of schooling ($L_{SY1}$) and compulsory attendance ($L_{SY2}$).\(^10\)

Let us inspect the replicate of SY’s results first (see Figure 4). The IV-based model specifications appear to lack empirical consistency and robustness relative to their OLS counterpart. $\beta^L$ 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 (2016).

---

\(^9\) 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 ($L_{AK1}$), second ($L_{AK2}$) and third ($L_{AK3}$) quarter of the year; see also Angrist and Krueger (1992).

\(^10\) As in AK the indicators compose of three dummies. $L_{SY11}$, $L_{SY12}$, and $L_{SY13}$ capture those with minimum of 7 or below, 8, and 9 or above required years of schooling and $L_{SY21}$, $L_{SY22}$, and $L_{SY23}$ 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.
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 $\beta^L$ whereas $\beta$ remains virtually invariant; see Figure 4 and Table 1 columns (2) and (4) of T1A and T1B. At the same time, the inclusion of interaction terms invalidates the claim of endogeneity if using $L_{SY2}$ indicators and leads to an insignificant $\beta^L$ estimate if using $L_{SY1}$ indicators. The sensitivity of IV estimates to regional factors is pointed out by SY and reiterated by Hogerheide and van Dijk (2006: Table 5). This raises the question whether $L_{SY}$ solely represent the CSL treatment; a potential case of measurement error.

---

**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 (left axis for IV estimates and right axis for OLS estimates). The bars indicate the 95% confidence interval. Source: SY, Table 1.

**Figure 4. OSL and 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 (left axis for IV estimates and right axis for OLS estimates). The bars indicate the 95% confidence interval. Source: SY, Table 1.

---

11 CSL indicators based on quarter of birth dummies face similar problems and Bound and Jaeger (2000) and Carneiro and Heckman (2002) show an entanglement of indicators with social status.
TABLE 1—SARGAN AND HAUSMAN TEST FOR INSTRUMENTS USED BY SY

| White males | T1A (ℒSY1) | T1B (ℒSY2) |
|-------------|------------|------------|
|             | aged 40-49 | aged 25-54 | aged 40-49 | aged 25-54 |
|             | (1)        | (2)        | (3)        | (4)        |
| β (OLS)²    | 0.073**    | 0.073**    | 0.063**    | 0.063**    |
| β² (2SLS)²  | 0.095**    | -0.020     | 0.097**    | -0.014     |
| Tests:      |            |            |            |            |
| Sargan²     | 0.99       | 4.65       | 17.99      | 7.51       |
| (p-value)   | (0.6888)   | (0.9777)   | (0.0001)   | (0.0234)   |
| Hausman     | 3.80       | 9.67       | 43.24      | 36.32      |
| (p-value)   | (0.0512)   | (0.0019)   | (0.0000)   | (0.0000)   |
| Fixed effects: |          |            |            |            |
| State of birth | Yes       | Yes        | Yes        | Yes        |
| Year of birth | Yes       | Yes        | Yes        | Yes        |
| Region x Yob | No        | Yes        | No         | Yes        |
| Additional controls: | None    | None       | Age quartic, census yr | Age quartic, census yr |

Notes: Sargan and Hausman added through replication. Source: SY Tables 1 and A2. * Robust and cluster adjusted standard errors are used. ** Significant at the 1 percent level. * Significant at the 5 percent level.

The insignificance and empirical inconsistency of β², identified in Figure 1 and Table 1, could also be caused by a negligible share of ‘compliers’ in the full sample; a point made by Oreopoulos (2006a) in the context of the CSL effect when using minimum years of schooling indicators. 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 is expected to be zero. 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 ℒSY2 indicators, roughly 4.11 per cent of the 1930-39 born cohort complies with the law. The share of compliers is even smaller for the later born cohort with 2.31 per cent. Using ℒSY2 indicators instead, the share of compliers is similarly small with 4.18 and 2.53 per cent in the 1930s and 1940s birth cohort respectively (see Table 1A, Appendix). Our rough estimates of complier shares are slightly lower than in Bolzern and Huber (2017), who report a complier share of 6-12 per cent for European countries based on comparison of mean potential outcomes using binary treatment and instrument variables.

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 localized treatment, i.e. treatment which is confined to specific complier groups; see Angrist et al. (1996) and Angrist and Imbens (1995). The estimates are thus interpreted as ‘local average treatment effect’

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.
(LATE) instead of ATE, e.g. see Angrist and Pischke (2009: Chapter 4), and also Heckman and Urzua (2009), Deaton (2009), Imbens (2010). A simpler way of examining the localness is to estimate the models using sub-sample data. Therefore, we replicate SY Tables 1 and A2 and AK Tables V and VI, using sub-samples to separate ‘always takers’ from ‘compliers’. 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 1A, Appendix). The aim of this experiment is to examine the empirical consistency of the two models under the localized treatment condition.

The IV model estimates are close to the OLS counterparts for most of the School sub-sample, while $\beta^L$ is insignificant for the Higher sub-sample, confirming our conjecture of zero ATE by CSL treatment for the latter sub-sample (Table 2). 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 $L_{SY1}$ indicators for the School sub-sample. Interestingly, the OLS estimates, while significant and positive throughout all sub-samples and cohorts, vary across sub-samples, with the ARTE decreasing for those attaining 13 to 15 years of education.

The results of the above experiments show a notable lack of consistency in the IV case. To formally compare the two model specifications, we draw on the aforementioned insight from machine learning theory; see Mukherjee et al. (2006) and Shalev-Shwartz et al. (2010). CV techniques separate 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 MSE in a k-fold CV, e.g. see Arlot and Celisse (2010), Zhang and Yang (2015). Figure 5 shows that the OLS-based model clearly outperforms the IV-based one in generalizability, stability and consistency. Remarkably, results for AK are close to or even worse than for SY, despite the finding of $\beta \approx \beta^L$ in AK.

---

13 The separation remains imperfect, as some always takers will be contained in the complier group.
14 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.
15 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.
TABLE 2—ESTIMATION OF $\beta$ VIA SUB-SAMPLING ON EDUCATIONAL ATTAINMENT

| Years of schooling | School | 7-12 | 13-15 | Higher | ≥ 13 | School | 7-12 | 13-15 | Higher | ≥ 13 |
|--------------------|--------|------|-------|--------|------|--------|------|-------|--------|------|
| SY                 |        |      |       |        |      |        |      |       |        |      |
| $\beta$            |        |      |       |        |      |        |      |       |        |      |
| [95% CI]           |        |      |       |        |      |        |      |       |        |      |
| $\beta^a$ (with $L_{OLS}$) | .0942** | .060* | -2.677 | -2.018 | .1008** | .2072** | .449 |        |        |      |
| [95% CI]           |        |      |       |        |      |        |      |       |        |      |
| 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) | 9.98 (0.3230) |        |      |
| $\beta^a$ (with $L_{OLS}$) | .1954** | 1.141* | -5.833 | -0.831 | .0836** | .671 | 0.529 |        |        |      |
| [95% CI]           |        |      |       |        |      |        |      |       |        |      |
| Sargan             | .3182 (0.2010) | .0193 (2.630) | -1.76 (5.91) | -1.34 (5.07) | 0.0499 (1.72) | -0.30 (1.64) | -1.20 (0.91) | 2.20 (0.009) |        |      |
| Hausman            | 24.6 (0.0000) | 1.86 (1.732) | 1.79 (1.807) | 0.02 (0.962) | 1.72 (1.986) | 0.32 (1.879) | 1.75 (1.162) | 0.11 (0.7419) |        |      |
| AK                 |        |      |       |        |      |        |      |       |        |      |
| $\beta$            |        |      |       |        |      |        |      |       |        |      |
| [95% CI]           |        |      |       |        |      |        |      |       |        |      |
| $\beta^a$ (with $L_{OLS}$) | .0655** | .0581** |       | .0442** | .0591** | .0701** | .074** | .053** | .045** |        |
| [95% CI]           |        |      |       |        |      |        |      |       |        |      |
| Sargan             | .3551 (0.078) | .0654 (0.061) | .0371 (0.155) | .0575 (0.067) | .1868 (0.0716) | .9715 (0.0.753) | .0105 (0.001) | .1093 (0.0416) |        |      |
| Hausman            | 24.6 (0.0000) | 24.2 (0.7172) | 33.4 (0.2633) | 30.1 (0.497) | 60.8 (0.0005) | 59.6 (0.0007) | 38.0 (0.1230) | 33.5 (0.2587) |        |      |

Notes: SY data follows model specification (1) Tables 1 and 2; 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 (.). * Standard errors cluster adjusted. ** Significant at the 1 percent level. * Significant at the 5 percent level.

Figure 5. Ratio of Average Mean Square CV Error of IV to OLS With Increasing K

Note: 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 one indicates a smaller MSE for OLS than for IV.

As expected, the MSE is decreasing 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 and the OLS-based model outperforms the IV-based one at all k, even though
the CV experiment presented here does not adjust for degrees of freedom.\footnote{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.} When decomposing the MSE into test bias and variance, we find little evidence of asymptotic bias in the OLS estimates (Figure 6). Instead, there are small but discernible decreases of the OLS bias while the IV bias increases with \( k \), suggesting over fitting in the IV-based model. Given the small bias, the large difference in the MSE between the two models clearly stems from a greater variance or instability of the IV model specification, putting the consistency claim of IV into question. While our findings are specific to the CSL case and the chose instruments, results by Young (2017) who re-evaluates 1,359 published IV regressions, suggest that the conclusion drawn from the CV exercise are the norm rather than an exception.

![Image of Figure 6: Average CV Bias with Increasing K](image)

**Note:** 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 \( k \) shown on the x-axis.

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 one in terms of generalizability, stability and consistency, regardless the choice of CSL indicators.

### 4. CSL Treatment in A Multiple Model Framework

Given the overwhelming rejection of \( s^L \) in favor of \( s \), we now probe further into the evidence of \( s \) as a valid conditional variable by rout of the chain model approach presented in Section 2. The chain model approach involves the estimation of the two outcome effects on earnings: (a) the ARTE effect, \( \beta_{yzZ} \), and (b) the CSL effect or ATE of the CSL via schooling, \( \beta_{yLs} \). We then
investigate the possible presence of a L-treatment induced parametric shift in the schooling effect, \( \beta_{y_{sL}Z} \neq \beta_{y_{sL}Z} \).

### 4.1. Sub Estimating the Schooling Effect: \( \beta_{y_{s}Z} \)

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 (Table 2). 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 also can align the ARTE estimates by AK and SY data respectively. This is achieved through firstly, unification 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. As for the experiments reported in Table 2, 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 \((\text{age}^2)\).\(^{17}\) Regional dummies for individual states are replaced by a single variable distinguishing between four regions for SY and nine regions for AK data \((\text{region})\). Considering a possible regional effect on school quality, variables capturing school quality \((\text{pupilt}, \text{term}, \text{reltwage})\) suggested by Card and Krueger (1992a; 1992b) are used by SY and included in our model as well.

A notable pattern of parameter inconstancy in Table 2 is the variation in the ARTE estimates with the level of education. The variation reflects ‘sheepskin effects’, which are well documented phenomena in the literature\(^{18}\) and clearly discernible in the AK and SY data; see Figure 1A and Table 1A, Appendix. A dummy/binary variable \((\text{uni})\) 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 3, alongside with those from the ‘Original’ models by SY and AK. We refer to our more parsimonious specifications as ‘Alternative’ in the table. An alignment of return to schooling estimates across datasets is achieved with the ‘Alternative’ model specification, which outperforms the ‘Original’ specification in terms of model fit. In particular, OLS estimates point to a constant \( \beta_{y_{s}Z} \) of

---

\(^{17}\) 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.

\(^{18}\) 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).
about 0.06 across data sets and cohorts. Our ARTE estimates are roughly in line with findings by Pischke and von Wachter (2008) and Acemoglu and Angrist (2001), who report estimates of 0.061 and 0.075 respectively.

Table 3—Parsimonious Specification of (9)

|         | Original SY |        |         | Original AK |        |         |         | Alternative SY |        |         |         | Alternative AK |
|---------|-------------|--------|---------|-------------|--------|---------|---------|----------------|--------|---------|---------|----------------|
| $\beta_{ys,z}$ | 1930-1939 | 1940-1949 | 1930-1939 | 1940-1949 | 1930-1939 | 1940-1949 | 1930-1939 | 1940-1949 | 1930-1939 | 1940-1949 | 1930-1939 | 1940-1949 |
| [95% CI] | [0.0751**] | [0.0622**] | [0.0519**] | [0.0630**] | [0.0600**] | [0.0643**] | [0.058 .062] | [0.063 .066] | [0.057 .059] | [0.064 .066] |
| AIC*    | 714262.9 | 1034376 | 594994.7 | 858645.2 | 705271.2 | 1018112 | 594343.4 | 858594.8 |
| Adj.-R² | 0.0119 | -0.0232 | 0.1745 | 0.1354 | 0.1217 | 0.0968 | 0.1761 | 0.1355 |
| Consist.| 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| $Z^2$   | age2, age3, age4, yob31, yob39, yob41, yob49, sob1-sob55 | ageq, ageq2, race, married, sma, neweng midatl, enocent, wnocent, soat, esocent, wsocent, mt, yr20-yr28 | age2, mar, emp, jail, handicap, pupil, term, reltwait, uni, region |

Notes: 1980 census, data for SY white male with positive weekly earnings, data for AK male with positive weekly earnings. * 95 per cent confidence interval based on cluster adjusted standard errors in SY data. * Akaike information criteria. * Entner et al. (2012) test for consistency. The row reports the correlation coefficient between $\epsilon_y$ in (9) and the residuals from the auxiliary regression with a value close to zero confirming consistency. Non-Gaussianity of the residuals was tested before and strongly supported by data. * See SY and AK for variable names. ** Significant at the 1 percent level. * Significant at the 5 percent level.

As discussed earlier and indicated by the findings in Table 3, 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 $\epsilon_y$ 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 $Z. 19$ If residuals of this auxiliary regression are non-Gaussian – Gaussian residuals are a rarity in large cross-sectional data sets – it is tested for statistically independent between $\epsilon_y$ 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 3. In all cases, consistency is strongly supported by the data, rejecting the presupposition of OVB underlying the IV treatment.

19 The auxiliary regression takes the form $s = \alpha_s + Z^\prime \beta_{ys,z} \epsilon_s + \epsilon_s$. If $\epsilon_s$ is non-Gaussian, statistical independence between $\epsilon_s$ and $\epsilon_y$ confirms consistency of $\beta_{ys,z}$ in (9).
4.2. Estimating the ATE of the CSL via Schooling: $\beta_{yLs}$

Given the potential measurement error in CSL indicators identified by SY and briefly discussed in Section 3, we conduct two simple experiments to further test the appropriateness of the indicator choice before continuing with the estimation of $\beta_{yLs}$. 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 subsample groups by educational attainment as in Table 2, with the expectation that $\hat{\beta}_{sl} \neq 0$ for School and $\hat{\beta}_{sl} = 0$ for Higher.

It is shown in Table 4 that, although $\hat{\beta}_{sl}$ tends to be larger for the School subsample 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 1A (Appendix), the share of those attaining less or equal the minimum years of schooling is halved in the later cohort.

| 1940-1949 | School $L_{yr1}$ | Higher $L_{yr2}$ | School $L_{yr1}$ | Higher $L_{yr2}$ | School $L_{AK}$ | Higher $L_{AK}$ |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1939      | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    |
| $\hat{\beta}_{sl}$ | 0.38* | 2.50 | 0.06 | 0.69 | 0.18 | 1.84 | -0.1** | -4.33 | -0.1** | -8.74 | 0.02 | 1.17 |
| $\hat{\beta}_{sl}$ | 0.35* | 2.49 | 0.07 | 0.91 | -0.02 | -0.23 | -0.05* | -1.98 | -0.1** | -8.43 | 0.05** | 3.10 |
| $\hat{\beta}_{sl}$ | 0.18 | 1.25 | 0.06 | 0.75 | 0.39** | 3.99 | -0.1** | -4.27 | -0.03* | -2.39 | -0.00 | -0.04 |

| 1940      | School $L_{yr1}$ | Higher $L_{yr2}$ | School $L_{yr1}$ | Higher $L_{yr2}$ | School $L_{AK}$ | Higher $L_{AK}$ |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1949      | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    | Coef. t-stat    |
| $\hat{\beta}_{sl}$ | 0.70** | 11.34 | 0.13** | 3.47 | 0.24** | 4.59 | -0.05 | -1.57 | -0.1** | -9.89 | 0.04** | 3.71 |
| $\hat{\beta}_{sl}$ | 0.69** | 10.61 | 0.25** | 7.63 | -0.03 | -0.57 | 0.03 | 1.00 | -0.1** | -7.82 | 0.06** | 5.09 |
| $\hat{\beta}_{sl}$ | 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. * Robust cluster adjusted standard errors. ** Robust standard errors. ** Significant at the 1 percent level. * Significant at the 5 percent level.

In a second step, we test whether the rule of intervention $\beta_{yLZS} = 0$ 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 5 that the condition $\beta_{yLZS} = 0$ is validated for SY’s $L_{SY}$ indicator across cohorts and also for AK’s $L_{AK}$ indicator for the early born cohort. But it is violated without exception if using $L_{SY2}$ as CSL indicator. Where conditional independence is rejected in Table 5, we have also failed to confirm $\hat{\beta}_{sl} = 0$ for the Higher sub-sample (see
Table 4) and rejected instrument validity (see Table 2). In cases like this, we should be cautious with the estimate of $\beta_{yLs}$ via the chain representation of (10).

**Table 5—Test for the Rule of Intervention $\beta_{yLZs} = 0$ Using (8) Extended by $Z$**

|          | SY*  |          | AK*  |
|----------|------|----------|------|
|          | 1930-39 | 1940-49 | 1930-39 | 1940-49 |
| $L_{y1}$ | -0.001 | .041**  | .007 | .041**  |
|          | (.0133) | (.0015) | (.0125) | (.0099) |
| $L_{y2}$ | .014 | .041**  | .007 | .046**  |
|          | (.0146) | (.0106) | (.0030) | (.0039) |
| $L_{y3}$ | .007 | .007*  | .008* | .01**  |
|          | (.003) | (.003) | (.003) | (.003) |
| $L_{s1}$ | .011 | .028**  | .026 | .052**  |
|          | (.0126) | (.0081) | (.0053) | (.0024) |
| $L_{s2}$ | .016 | .047**  | .020 | .004  |
|          | (.0142) | (.0085) | (.0030) | (.0039) |
| $L_{s3}$ | .033 | .063**  | .033** | .017** |
|          | (.0122) | (.0113) | (.0124) | (.0038) |
| 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 (.). a Robust cluster adjusted standard errors. b Robust standard errors. ** Significant at the 1 percent level. * Significant at the 5 percent level.**

Table 6 provides $\beta_{yLs}$ estimated via (10). Where conditional independence was verified, the chain approximation yields significant ATE estimates that confirm our expectation of a larger effect of schooling on earnings than that of the CSL, i.e. $\beta_{yLs} \gg \beta_{yLs}$. Direct ATE estimates $\beta_{yL}$ obtained via (7’) exceed estimates obtained via chain approximation for the later born cohort (see Table 2A, Appendix). 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 $\beta_{sL}$ in Table 4. The estimated ATE almost doubles for the later born cohort from 1–3 to 3–5 per cent using $L_{SY1}$ 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 $L_{AK1}$, $L_{AK2}$ and $L_{AK3}$ 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.

Our ATE estimates are in rough agreement with findings reported in the literature. For instance, Pischke and von Wachter (2008) report the ATE of CSL in Germany to be between 0.012 and 0.017 for different datasets. Results presented by Oreopoulos (2006b) for Canada suggest a larger ATE between 0.031 and 0.107, which is similar to our estimates using $L_{SY1}$. Acemoglu and Angrist (2001) report the ATE of CSL to be between 0.008 and 0.009. A
moderate positive impact of CSL on schooling was further confirmed by Lleras-Muney (2002), Oreopoulos (2006a) and Goldin and Katz (2011).

### Table 6—Estimated ATE of CSL, $\beta_{y',L}$, Using Chain Model (9)

|            | SY        |            | AK        |
|------------|-----------|------------|-----------|
|            | 1930-39   | 1940-49    | 1930-39   | 1940-49   |
|            | Full      | School     | Full      | School     |
| $\gamma_1$ | .025**    | .011*      | .022*     | .011*      |
|            | [12.5]    | [2.56]     | [6.25]    | [3.40]     |
| $\gamma_2$ | .011*     | .016*      | .016**    | .016**     |
|            | [20.8]    | [20.8]     | [6.26]    | [88.4]     |
| $\gamma_3$ | .005      | .011*      | .023**    | .027**     |
|            | [0.97]    | [0.05]     | [35.4]    | [0.39]     |
| $\gamma_4$ | .001      | .004       | .028**    | .014**     |
|            | [107]     | [0.32]     | [107]     | [0.32]     |
|           | .004      | .002       | .002*     | .002*      |
|            | [0.44]    | [0.12]     | [37.8]    | [11.4]     |
|           | .012      | .022*      | .028**    | .014**     |
|            | [1.56]    | [20.3]     | [37.8]    | [4.62]     |
|           | .011      | .002       | .002*     | .002*      |
|            | [15.9]    | [0.12]     | [11.4]    | [5.68]     |
|           | .023**    | .027**     | .014**    | .002*      |
|            | [20.3]    | [0.12]     | [37.8]    | [4.62]     |
|           | .002*     | .002*      | .002*     | .002*      |
|            | [15.1]    | [6.00]     | [11.4]    | [5.68]     |

**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 $\beta_{y,L}$ and $\beta_{y,SL}$ estimates respectively. Significance of $\beta_{y,L}$ based on $\chi^2$ statistics estimated following Wee (1999), reported in [1].

* Robust cluster adjusted standard errors. ** Robust standard errors. * Significant at the 1 percent level. ** Significant at the 5 percent level.

#### 4.3. Testing for a CSL Induced Shift in the Schooling Effect: $\beta_{ySL,Z} \neq \beta_{yS,L,Z}$

If the introduction of the CSL has altered the ability composition of workers, the post treatment schooling effect, $\beta_{ySL,Z}$, might differ from the schooling effect before treatment, $\beta_{yS,L,Z}$, namely $\beta_{ySL,Z} \neq \beta_{yS,L,Z}$. To investigate if such a parameter shift has occurred, breakpoint 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 1A, Appendix) — the ability composition is unlikely to render a significant parameter shift using the full sample. As before, sub-sample division is used to minimizes 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 7, despite the careful sub-sample division, the null hypothesis of no break point cannot be rejected at the 5 per cent level in all experimental settings.

---

20 Although Table 1A (Appendix) is based on SY data, the patterns are likely to be identical for the AK data.
For the SY indicators the sub-sample division is slightly more complicated. $L_{SY}$ indicators capture the minimum years of schooling and school attendance required by the respective state’s labor 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 minimize 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 minimizing the number of those defecting in the group under the given information.

As can be seen from Table 8, using $L_{SY1}$ indicators, the break-point Chow tests provide no evidence for a treatment induced parametric shift in the ARTE parameter at the 1 per cent significance level and some evidence at 5 per cent significance level for the School Binding sub-sample. Using $L_{SY2}$ 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 per cent level for the Full sample and School sub-sample. However, given measurement errors identified for the $L_{SY2}$ indicators in Table 4, the evidence is too weak to conclude on a parametric shift.

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 Tables 2 and 3).
**TABLE 8—TEST FOR LINDEMAN INDUCED PARAMETRIC SHIFT IN $\beta_{\gamma LZ}$**

|                | Full | School (7-12y) | Higher (13-15y) | School Binding* |
|----------------|------|----------------|-----------------|-----------------|
|                | $L_{S1}$ | $L_{S2}$ | $L_{S1}$ | $L_{S2}$ | $L_{S1}$ | $L_{S2}$ |
| Chow Test (Treated-Full)* | 0.01 | 1.80 | 0.01 | 0.06 | 0.46 | 0.09 | 1.26 | 2.59 |
|                  | (0.9172) | (0.1797) | (0.9236) | (0.8056) | (0.4990) | (0.7693) | (0.2607) | (0.1076) |
| Chow Test (Untreated-Full)* | 0.02 | 0.37 | 0.03 | 0.10 | 0.38 | 0.20 | 4.45* | 0.19 |
|                  | (0.8848) | (0.5417) | (0.8574) | (0.7556) | (0.5395) | (0.6569) | (0.0350) | (0.6596) |

|                | Full | School (7-12y) | Higher (13-15y) | School Binding* |
|----------------|------|----------------|-----------------|-----------------|
|                | $L_{S1}$ | $L_{S2}$ | $L_{S1}$ | $L_{S2}$ | $L_{S1}$ | $L_{S2}$ |
| Chow Test (Treated-Full)* | 2.24 | 2.28 | 1.49 | 1.96 | 0.02 | 0.00 | 0.00 | 14.05** |
|                  | (0.1342) | (0.1312) | (0.2220) | (0.1617) | (0.8847) | (0.9896) | (0.9580) | (0.0002) |
| Chow Test (Untreated-Full)* | 5.04 | 4.23* | 3.25 | 6.06* | 0.05 | 0.02 | 0.26 | 2.41 |
|                  | (0.0248) | (0.0396) | (0.0713) | (0.0138) | (0.8195) | (0.8966) | (0.6115) | (0.1206) |

**Notes:** 1980 census, white male with positive weekly earnings. P-values in (.). * 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 zero. 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 zero. 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.

5. **What have we learnt?**

Angrist and Pischke (2015: 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. Our analysis shows that the failure lies in model design rather than choice of IVs, and that the failure is caused by nescience of the causal model alternation nature of the IV approach. A re-assessment of the approach focusing on its effects on causal model design in Section 2 leads us to identify several inconsistencies. The IV treatment was motivated by the desire for consistent estimation of model (1) out of concerns over the presence of OVB, measurement error and/or self-selection bias linked to innate ability when (1) is applied to data. However, an extension of model (1), by taking into careful consideration the causal implications of these concerns via DAGs in Section 2, shows that the relevance of these concerns crucially depends on the specified parameter of interest subsequent to the choice of the key conditional variable.

If the parameter of interest is the ATE of the CSL via schooling, concerns regarding selection bias and measurement error in the schooling variable are irrelevant, since $s$ merely acts as a mediator for $L$. As evident from Tables 4–6, measurement error in CSL indicators is a major concern. CSL indicators fail to consistently target school goers, fail the rule of intervention and
result in non-robust ATE estimates. These weaknesses in the CSL indicators call for a more careful indicator selection; a concern raised repeatedly in the literature. If the parameter of interest is the ARTE, the effect of \( L \) is irrelevant, unless its moderating effect on \( s \) induces a parametric shift. We find little evidence of such a shift using the CSL indicators of AK and SY (Tables 7 and 8).

The IV approach, in contrast, tries apparently to uphold model (1) through compounding \( L \) with \( s \). This effectively rejects \( s \) as the valid conditional variable by replacing it with the IV-compounded \( s^L \) in model (3), thus inadvertently rejecting (1). While \( \beta^L \) is taken as a consistent estimator of (1), its causal interpretation creates unceasing row. Theoretically, the non-unique nature of \( s^L \) deems the non-existence of a unique \( \beta^L \). Empirically, compounding \( L \) with \( s \) effectively reweights the samples indirectly, reducing the ATE estimates to a set of LATE estimates which are conceptually not unifiable. Such ‘localised’ interpretation, unfortunately, makes the empirical findings drift further away from model (1).

Once the choice between the IV and OLS is viewed through the lens of choosing rival conditional variables, or rival model designs, empirical tests are called for. Repeated k-fold CV experiments reported in Section 3 yield no trace of IV estimates converging with increasing sample sizes while OLS estimates show no trace of being asymptotically biased, whether the samples are ‘localised’ or not. In contrast, the notably stable OLS estimates and their higher predictive accuracy, or higher generalizability, serve as strong support for \( s \) being more conditionally valid than \( s^L \). The evidence shows it a grave misperception against the OLS as being inconsistent or uninteresting while the crux lies in a priori causal model misspecification.

The CV experiments help us deselect model (3) in favour of (1); see Arlot and Celisse (2010). Moreover, they also teach us the primary importance of building and selecting data-consistent models to minimize empirical risk. It is a problematic strategy to go for consistent estimators before model selection is accomplished; see an insightful discussion on learnability versus consistency by Shalev-Shwartz and Ben-David (2014: Chapter 7).

The selection step entails a causally explicit model design. In the event that the causal effect of the key conditional variable is compounded by another variable, a multivariate model specification is essential and a clear differentiation between moderator and mediator effects therein is imperative. That way, concerns arising over measurement error, selection bias and OVB can be carefully assessed at the level of individual parameters as demonstrated in Section 4. For instance, our model design reveals that the greater risk of OVB with respect to the ARTE
estimate comes from inadequate specification of $Z$ rather than the possible omission of $L$, opposite to what underlies the IV-based research design.

Our results clearly demonstrate the importance of appropriate model design through statistical data learning. The consideration of consistent estimation should not be called for until a faithful translation of causal postulates into data-consistent statistical models is established.

References

Abadie, A. and Cattaneo, M.D. (2018) Econometric Methods for Program Evaluation, Annual Review of Economics, forthcoming.

Acemoglu, D., J. D. Angrist (2001), How Large Are Human Externalities? Evidence from Compulsory Schooling Laws, NBER Macroeconomics Annual 15: 9–74.

Angrist, J. D. (1995), The Economic Returns to Schooling in the West Bank and Gaza Strip, The American Economic Review 85(5): 1065–1087.

Angrist, J. D., G. 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(430): 431–442.

Angrist, J. D., G. W. Imbens, D. B. Rubin (1996), Identification of Causal Effects Using Instrumental Variables, Journal of the American Statistical Association 91(434): 444–455.

Angrist, J. D., A. B. Krueger (1991), Does Compulsory School Attendance Affect Schooling and Earnings? Quarterly Journal of Economics 106(4): 979–1014.

— (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(418): 328–336.

Angrist, J. D., J.-S. Pischke (2015), Mastering 'Metrics: The Path from Cause to Effect, Princeton, New Jersey and Woodstock, Oxfordshire: Princeton University Press.

— (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton, New Jersey and Woodstock, Oxfordshire: Princeton University Press.

Arlot, S., A. Celisse (2010), A Survey of Cross-Validation Procedures for Model Selection, Statistical Surveys 4: 40–79.

Athey, S., G. W. Imbens (2015), Machine Learning Methods for Estimating Heterogeneous Causal Effects, Mimeo, Stanford University, Stanford, CA, USA.

Bolzern, B., M. Huber (2017), Testing the Validity of the Compulsory Schooling Law Instrument. Economics Letters 159: 23–27.
Bound, J., D. 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, D. (2001), Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems, Econometrica 69(5): 1127–60.

Card, D., A. 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): 1–40.

Card, D., A. B. Krueger (1992b), School Quality and Black-White Relative Earnings: A Direct Assessment, Quarterly Journal of Economics 107(1): 151–200.

Carneiro, P., J. J. Heckman (2002), The Evidence on Credit Constraints in Post-Secondary Schooling, The Economic Journal 112: 989–1018.

Clark, D., P. Martorell (2014), The Signaling Value of a High School Diploma, Journal of Political Economy 122(2): 282–318.

Cox, D. R., N. Wermuth (1996), Multivariate Dependencies – Model, Analysis and Interpretation, Chapman & Hall.

Cox, D. R., N. Wermuth (2004), Causality: A Statistical View, International Statistical Review 72(3): 285–305.

Deaton, A. (2009), Instruments of Development: Randomization In The Tropics, And The Search For The Elusive Keys To Economic Development, NBER Working Paper Series, No. 14690.

Deaton, A., N. Cartwright (2016), Understanding and Misunderstanding Randomized Controlled Trials, NBER Working Paper Series, No. 22595.

Durbin, J. (1954), Errors in Variables, Review of the International Statistical Institute 22(1): 23–32.

Engle, R. F., D. F. Hendry, J.-F. Richard (1983), Exogeneity, Econometrica 51(2): 277–304.

Entner, D., P. O. Hoyer, P. Spirtes (2012), Statistical Test for Consistent Estimation of Causal Effects in Linear Non-Gaussian Models, Appearing in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Goldin, C. (1998), America's graduation from high school: The evolution and spread of secondary schooling in the twentieth century, Journal of Economic History 58(2): 345–374.

Goldin, C., L. F. Katz (2000), Education and Income in the Early 20th Century: Evidence from the Prairies, The Journal of Economic History 60(3): 782–818.

Goldin, C., L. F. Katz (2007), Long-Run Changes in the U.S. Wage Structure: Narrowing,
Goldin, C., L. 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, by D. Costa and N. Lamoreaux, 275–310. Chicago: University of Chicago Press.

Harmon, C., H. Oosterbeek, I. Walker (2003), The Returns to Education: Microeconomics, Journal of Economic Surveys 17(2): 115–155.

Heckman, J. J., S. Urzua (2009), Comparing IV with Structural Models: What Simple IV Can and Cannot Identify, NBER Working Paper Series, No. 14706.

Hoogerheide, L., H. K. van Dijk (2006), A Reconsideration of the Angrist-Krueger Analysis on Returns to Education, Econometric Institute Report EI 2006-15.

Imbens, G. W. (2010), Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009), Journal of Economic Literature 48(2): 399–423.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, G. W. Imbens (2015), Identification and Inference with Many Invalid Instruments, Journal of Business & Economic Statistics 33(4): 474–484.

Lleras-Muney, A. (2002), Were Compulsory Attendance and Child Labor Laws Effective? An Analysis from 1915 to 1939, The Journal of Law and Economics 45(2).

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling, L. Sanbonmatsu (2013), Long-Term Neighborhood Effect on Low-Income Families: Evidence from Moving to Opportunity, American Economic Review 103(3): 226–231.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling, L. Sanbonmatsu (2012), Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults, Science 337(6101): 1505–1510.

Mukherjee, S., P. Niyogi, T. Poggio, R. Rifkin (2006), Learning Theory: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization, Advances in Computational Mathematics 25: 161–193.

Murphy, K. M., F. Welch (1992), The Structure of Wages, The Quarterly Journal of Economics 107(1): 285–326.

Oreopoulos, P. (2006a), Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter, American Economic Review 96(1): 152–175.

Oreopoulos, P. (2006b), The Compelling Effects of Compulsory Schooling: Evidence from Canada, The Canadian Journal of Economics 39(1): 22-52.

Pearl, J. (2009), Causality: Models, Reasoning, and Inference (2nd edition). Cambridge, US:
Cambridge University Press.
Pischke, J.-S., T. von Wachter (2008), Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation, The Review of Economics and Statistics 90(3): 592–598.
Qin, D. (2019), Let’s Take the Bias Out of Econometrics, Journal of Economic Methodology, 26(June) forthcoming.
Qin, D. (2015), Resurgence of the Endogeneity-Backed Instrumental Variable Methods, Economics: The Open-Access, Open-Assessment E-Journal 9(7): 1–35.
Shalev-Shwartz, S., S. Ben-David (2014), Understanding Machine Learning: From Theory to Algorithms, Cambridge: Cambridge University Press.
Shalev-Shwartz, S., O. Shamir, N. Srebro, K. Sridharan (2010), Learnability, Stability and Uniform Consistency, Journal of Machine Learning Research 11: 2635–2670.
Stephens Jr., M., D.-Y. Yang (2014), Compulsory Education and the Benefits of Schooling, American Economic Review 104(6): 1777–1792.
Stock, J. H., J. H. Wright, M. Yogo (2002), A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments, Journal of Business & Economic Statistics 20(4): 518–529.
Trostel, P. A. (2005), Nonlinearity in the Return to Education, Journal of Applied Economics 8(1): 191–202.
Weesie, J. (1999), Seemingly Unrelated Estimation and The Cluster-adjusted Sandwich Estimator, Stata Technical Bulletin 52: 34–47. Reprinted in Stata Technical Bulletin Reprints 9: 231–248.
Wermuth, N., D. R. Cox (2011), Graphical Markov Models: Overview, in Wright, J. ed. International Encyclopedia of Social and Behavioral Sciences, 2nd ed. 10, Elsevier: 341–50.
Young, Alwyn. (2017) Consistency without Inference: Instrumental Variables in Practical Application, Mimeo, London School of Economics, London, UK.
Zhang, K., M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, C. Glymour (2017), Causal Discovery in the Presence of Measurement Error: Identifiability Conditions, Presented at UAI 2017 Workshop on Causality.
Zhang, Y., Y. Yuhong (2015). Cross-Validation for Selecting a Model Selection Procedure, Journal of Econometrics 187: 95-112.

Acknowledgments

This research was partly funded by an internal research fund from the Faculty of Law and Social Sciences, SOAS, University of London.
Appendix

a. 1930-1939 SY

b. 1930-1939 AK

c. 1940-1949 SY

d. 1940-1949 AK

**FIGURE 1A. YEARS FOR SCHOOLING DENSITY**

*Notes:* AK education variable is capped at 17 years of schooling for comparability between the AK and SY datasets.
### Table 1A—Composition of CSL Compliers, Defectors and Always Takers for $\mathcal{L}_{SY}$

| Composition of CSL Compliers for $\mathcal{L}_{SY1}$ | Years of Schooling | Untreated |
|--------------------------------------------------|-------------------|-----------|
| 1930-1939                                        |                   |           |
| Equal                                            | 1.37%             | 5.48%     |
| Less                                             | 2.09%             | 4.04%     |
| More                                              | 96.45%            | 90.57%    |
| N                                                | 54,992            | 116,112   |
|                                                   | 193,730           | 366,381   |
|                                                   | N (% of total)     | 1,547     |
|                                                   | 1939               | 0.42%     |
|                                                   | Less               | 8.69%     |
|                                                   | 96.86%            | 92.22%    |
|                                                   | N                  | 86,252    |
|                                                   | 112,481           | 337,970   |
|                                                   | N (% of total)     | 12,158    |

### Table 1B—Composition of CSL Compliers for $\mathcal{L}_{SY2}$

| Composition of CSL Compliers for $\mathcal{L}_{SY2}$ | Years of Schooling | Untreated |
|--------------------------------------------------|-------------------|-----------|
| 1930-1939                                        |                   |           |
| Equal                                            | 5.23%             | 4.11%     |
| Less                                             | 3.65%             | 10.67%    |
| More                                              | 91.20%            | 85.22%    |
| N                                                | 116,797           | 179,705   |
|                                                   | 36,489            | 366,381   |
|                                                   | N (% of total)     | 33,390    |
|                                                   | 1939               | 0.11%     |
|                                                   | Less               | 5.64%     |
|                                                   | 96.83%            | 91.36%    |
|                                                   | N                  | 131,875   |
|                                                   | 297,498           | 548,870   |
|                                                   | N (% of total)     | 37,016    |

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.

### Table 2A—Estimated ATE of the CSL $\mathcal{B}_{1c}$ Via (7')

|                           | SY*  | AK*  |
|---------------------------|------|------|
|                           | Full | School | Full | School | Full | School | Full | School |
| $\mathcal{L}_{SY1}$       | 0.033| 0.014 | 0.032| 0.017 | 0.121| 0.095 | 0.014| 0.008 |
| $\mathcal{L}_{SY2}$       | 0.077| 0.065 | 0.053| 0.041 | 0.617| 0.535 | -0.423| -0.302 |
| $\mathcal{L}_{FY}$        | (1.74)| (2.87) | (0.65) | (3.80) | (5.61) | (2.41) | (2.36) | (0.64) |
| $\mathcal{L}_{FY1}$       | 0.010| 0.22  | 0.004| 0.024 | 0.145| 0.050 | 0.119| 0.060 |
| $\mathcal{L}_{FY2}$       | -0.092| (1.48) | (0.02) | (1.69) | (5.74) | (2.28) | (6.60) | (3.40) |
| $\mathcal{L}_{FY}$        | -0.287| (3.92) | (0.92) | (3.64) | (5.40) | (3.40) | (8.92) | (3.40) |
| $\mathcal{L}_{FY1}$       | 0.005| 0.106 | 0.095 | 0.105 | 0.125 | 0.005 | -0.005 | 0.010 |
| $\mathcal{L}_{FY2}$       | -0.006| (4.85) | (0.31) | (6.49) | (5.75) | (3.88) | (3.90) | (6.03) |
| $\mathcal{L}_{FY}$        | (0.17) | (1.24) | (4.05) | (4.48) | (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 (). * Robust cluster adjusted standard errors. ** Robust standard errors. ** Significant at the 1 percent level. * Significant at the 5 percent level.

Electronic copy available at: https://ssrn.com/abstract=3282990