Empirical Economics (2023) 65:2979–2999
https://doi.org/10.1007/s00181-023-02441-7

REPLICATION STUDY

When to use matching and weighting or regression in instrumental variable estimation? Evidence from college proximity and returns to college

Stefan Tübbicke[1]

Received: 23 May 2022 / Accepted: 18 May 2023 / Published online: 8 June 2023

Abstract

Standard two-stage least squares (2SLS) regression remains dominant in instrumental variables estimation of causal effects even though the literature has shown that 2SLS may be inconsistent when effects are heterogenous and the instrument is only valid when conditioning on covariates. To show that this is not merely a hypothetical threat, this paper re-estimates the returns to college using college proximity as an instrument based on the data from Card (Aspects of labour market behavior: essays in honour of John Vanderkamp, University of Toronto Press, Toronto, 1995). The results show that 2SLS yields systematically larger estimates of the returns to college than more flexible estimators based on the instrument propensity score. In the full sample, differences amount to about 50 to 100%. This is due to the implicit conditional-variance weighting performed by 2SLS. Moreover, in line with the theoretical prediction by Sloczynski (When should we (not) interpret linear IV estimands as LATE? IZA discussion papers 14349, Institute of Labor Economics (IZA), 2021), findings suggest that the impact of the conditional-variance weighting is larger when instrument groups are not roughly the same size. Thus, it is advised to use 2SLS with caution and use estimators based on the instrument propensity score instead when groups are of different size and covariates are predictive of the instrument.

Keywords Instrumental variables · Semi- and nonparametric methods · Returns to education

JEL Classification C14 · C26 · I26

✉ Stefan Tübbicke
stefan.tuebbicke@iab.de

[1] Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany

1 Introduction

The number of available methods for causal inference has seen enormous growth in the last three decades (see Abadie and Cattaneo 2018, for a recent overview). Although progress has been tremendous, applied research does not always keep up. On the one hand, flexible semi- and non-parametric methods based on the propensity score (PS) are widely applied when estimating causal effects under the selection-on-observables assumption (e.g., see Austin and Stuart 2015; Thoemmes and Kim 2011). On the other hand, much of the literature using instrumental variable (IV) estimation to overcome bias due to unobserved factors relies on two-stage least squares (2SLS). This is despite the fact that 2SLS may yield inconsistent estimates of treatment effects when effects are heterogenous and covariates are predictive of the instrument (Abadie 2003).

This paper concentrates on the most common case in which the researcher aims to estimate causal effects of some treatment using a single binary IV. First, the paper reviews available results from the literature on implications of using standard linear-in-covariates 2SLS estimation under effect heterogeneity. Most importantly for this paper, the literature shows that 2SLS yields a ratio of conditional variance-weighted average of covariate-specific effects. If effects are indeed heterogenous and related to the PS, then 2SLS yields inconsistent estimates. Estimators based on the PS provide a consistent (Frölich 2007), readily-available and intuitive alternative. Hence, the paper briefly describes some basic IV estimators using the PS as well as the novel efficient covariate balancing approach by Heiler (2021). By re-estimating the returns to college using these approaches—exploiting college proximity as an instrument using the data by Card (1995)—the paper shows that the threat of obtaining inconsistent estimates when using 2SLS is not merely hypothetical. 2SLS yields systematically larger effect estimates than more flexible estimators based on the PS. Further inspection shows that this difference is mainly due to the implicit conditional-variance weighting performed by 2SLS.

This case study has been widely used to teach Economics students around the world about the use of IV methods to overcome bias due to unobserved confounders as well as the importance of effect heterogeneity. Moreover, the case study has been widely used in a variety of papers, see for example Tan (2006), Huber and Mellace (2015), Kitagawa (2015), Mourifié and Wan (2017), Andresen and Huber (2021), Sloczynski (2021), Sloczynski et al. (2022) and Blandhol et al. (2022). Most of these papers are concerned with instrument validity, an issue that is discussed but not of main interest in this paper. The only study known to the author that compares parametric estimators of the returns to college with more flexible estimators for this case study is Sloczynski et al. (2022). They too find sizable gaps in estimates. However, in contrast to this paper, they do not offer an explanation for this phenomenon.

The remainder of this paper is organized as follows: Sect. 2 reviews identification and estimation using IVs, Sect. 3 applies 2SLS and comparison methods based on the PS to the data. Section 4 concludes.

2 Identification and estimation using Instrumental variables

Assume we have an i.i.d. sample for i = 1,…,N units, where for each unit we observe some exogenous characteristics X_i, a binary treatment variable D_i, an outcome Y_i and a single binary instrument Z_i. Furthermore, assume that there is an unobserved confounder U_i that has an impact both the treatment variable D_i and the outcome Y_i. In the language of classical least squares regression, this creates an omitted variable bias and the selection-on-observables assumption fails (Wooldridge 2010, Chap. 4). To stick to the returns-to-college example used throughout this paper, conditioning on observed characteristics such as labor market experience or region of residence is insufficient to remove bias from standard regression or matching estimates of the effects of college attendance on wages if unobserved ability has an impact on the college decision and labor market earnings (Blackburn and Neumark 1993).

Under such circumstances, one can use IV techniques to estimate causal effects by exploiting variation in the treatment variable D_i through the instrument Z_i. When effects are heterogenous, IV methods identify local treatment effects, i.e. average effects for specific sub-populations influenced by the instrument. For this identification result to hold, the instrument needs to be exogenous, i.e. the instrument has to be as good as randomly assigned after conditioning on covariates and there must not be a direct effect of the instrument on the outcome. Moreover, the instrument must influence the treatment decision in a monotonous way. Imbens and Angrist (1994) introduce what Sloczynski (2021) calls "strong monotonicity", which is the assumption that the instrument weakly increases or decreases the treatment probability for everyone. Under this assumption, IV methods identify the local average treatment effect (LATE, Imbens and Angrist 1994), also called the complier average causal effect (CACE), i.e. the average treatment effect of individuals who act in line with the instrument. If defiers, i.e. individuals who act in the opposite direction of compliers, exist, and one is willing to assume that the sign of the effect of the instrument on treatment is determined solely by covariates ("weak monotonicity", Sloczynski 2021), the CACE may be recovered by averaging effects for individuals with covariate values estimated to behave in the direction of compliers. Moreover, a more general effect, the mover average causal effect (Kolésar 2013), i.e. the average treatment effect for compliers and defiers, is identified.

Using the standard potential outcomes framework (Imbens and Angrist 1994; Rubin 1974), define $D_i(1)$and $D_i(0)$ as the potential treatment states if the unit was assigned $Z_i = 1$ or $Z_i = 0$. If the instrument indeed has no direct impact on the outcome, one may write potential outcomes as $Y_i(d_i)$, with $Y_i(1)$and $Y_i(0)$ being the outcomes that would be observed under treatment and without. Assuming the instrument raises the chance of receiving treatment on average, the strong monotonicity assumption implies that for compliers $D_i(1) > D_i(0)$, i.e. they receive treatment if assigned $Z_i = 1$and they do not if assigned $Z_i = 0$. Based on these definitions and assumptions, the standard CACE can be written as

$$\Delta^{CACE} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]. \quad (1)$$

The MACE is defined as $\Delta^{MACE} = E[Y_i(1) - Y_i(0)|D_i(1) \neq D_i(0)]$ and can be recovered by using a reordered instrument Z_i^R, i.e. an adapted instrument which is reversed for defiers, defined as[1]

$$Z_i^R = Z_i I\Big(\delta^D(X_i) \geq 0\Big) + (1 - Z_i) I\Big(\delta^D(X_i) < 0\Big), \tag{2}$$

where $\delta^D(X_i) = E[D_i(1) - D_i(0)|X_i]$ is the covariate-specific average effect of the instrument on the treatment decision and $I(\cdot)$ is the indicator function (Sloczynski 2021).

For the empirical analysis of the paper, the exogeneity assumption is assumed to hold. Moreover, it is assumed that at least the weak monotonicity assumption holds as well. While the failure of the monotonicity assumption makes the interpretation of estimands difficult if not impossible, differences in estimates of these quantities are still interesting to inspect in order to understand the estimators' behavior under effect heterogeneity.

For the following exposition of estimation methods, assume that strong monotonicity holds.[2] While Frölich (2007) shows that the CACE is non-parametrically identified under exogeneity and strong monotonicity, most applied research still uses 2SLS to estimate effects using an IV. Typically, researchers model the outcome and treatment equations as linear functions of the instrument and covariates. That is, they build regression models that look something like

$$Y_i = \alpha_Y + \beta_Y' X_i + \gamma^Y Z_i + \varepsilon_i^Y \tag{3}$$

$$D_i = \alpha_D + \beta_D' X_i + \gamma^D Z_i + \varepsilon_i^D, \tag{4}$$

where it is (implicitly) assumed that slope-coefficients are constant and that ε_i^Y and ε_i^D are well-behaved error terms. The corresponding 2SLS estimator can be written as $\widehat{\Delta}_{2SLS} = \widehat{\gamma}^Y / \widehat{\gamma}^D$, i.e. the ratio of the reduced form (3) OLS coefficient $\widehat{\gamma}^Y$ and the first stage (4) OLS coefficient $\widehat{\gamma}^D$ on Z_i. Under effect heterogeneity and standard regularity conditions, Sloczynski (2021) shows that $\widehat{\Delta}_{2SLS}$ converges to[3]

$$\text{plim}\widehat{\Delta}_{2SLS} = \frac{E\big[\delta^Y(X_i)\delta^D(X_i)Var(Z_i|X_i)\big]}{E\big[\delta^D(X_i)Var(Z_i|X_i)\big]}, \tag{5}$$

where $\delta^Y(X_i) = E[Y_i(1) - Y_i(0)|X_i, D_i(1) > D_i(0)]$is the average covariate-specific effect of treatment on the outcome for compliers. Hence, 2SLS yields a

[1] As noted by Sloczynski (2021), this requires the estimation of $\delta^D(X_i)$

[2] If Z_i^R were known, all results presented would also hold for the estimation of the MACE. However, as it is unknown how the estimation of Z_i^R affects the behavior of estimators and deriving such results is beyond the scope of this paper, the author does not further discuss estimators based on the reordered instrument in this part. Nonetheless, Sect. 3 applies this methodology in order to provide evidence that a failure of the strong monotonicity does not drive differences between 2SLS and PS-bases estimators.

[3] This formula also follows immediately by combining results from Angrist (1998) on probability limits for OLS regressions and the continuous mapping theorem, as 2SLS is simply a ratio of two OLS coefficients.

conditional-variance weighted average of covariate-specific effects for compliers. As $Var(Z_i|X_i) = P(Z_i = 1|X_i)(1-P(Z_i = 1|X_i)$, weights attain a maximum when the PS $P(Z_i = 1|X_i) = 0.5$ (e.g., see Angrist and Pischke 2008). An important but typically under-appreciated consequence of this weighting is that $\text{plim}\widehat{\Delta}_{2SLS} \neq \Delta^{CACE}$ if $\delta^Y(X_i)$ and $\delta^D(X_i)$ depend on the PS. Depending on the correlation structure at hand, this may lead to substantial inconsistencies when using 2SLS.

As an alternative, this paper considers IV estimators of Δ^{CACE} based on the PS as derived by Frölich (2007) as well as a recent extension by Heiler (2021).[4] These estimators do not restrict effect heterogeneity. As a consequence, they are consistent even when effects are not homogenous (Frölich 2007, and Heiler, 2021).

The IV-matching estimator based on the PS pairs up each unit from the groups defined by the instrument with one, multiple or weighted averages of units from the opposite group based on the PS in order to infer the missing counterfactuals. Let $\widehat{Y_i(1)}$ and $\widehat{D_i(1)}$ denote the estimated counterfactuals for the outcome and the treatment variable if the unit was assigned $Z_i = 0$ as obtained by matching. Analogously, let $\widehat{Y_i(0)}$ and $\widehat{D_i(0)}$ be the estimated counterfactual outcome and treatment variable if the unit was assigned $Z_i = 1$. Based on this definition, the IV-matching estimator can be written as

$$\widehat{\delta}_{MAT} = \frac{\sum_{i=1}^{N} Z_i\left(Y_i - \widehat{Y_i(0)}\right) + (1 - Z_i)(\widehat{Y_i(1)} - Y_i)}{\sum_{i=1}^{N} Z_i\left(D_i - \widehat{D_i(0)}\right) + (1 - Z_i)(\widehat{D_i(1)} - D_i)}. \tag{6}$$

To estimate the PS, a standard logit regression is used. Moreover, kernel matching (KM) is employed as it has been shown to be among the top-performing PS-based matching methods in several simulation studies under the selection-on-observables paradigm (e.g., see Frölich 2004; Busso et al. 2014). More specifically, the matching procedure is implemented using an Epanechnikov kernel with a bandwidth chosen via weighted cross-validation (Galdo et al. 2008). To avoid extrapolation, common support is imposed via the min-max criterion by Dehejia and Wahba (1999) as is standard in the PS-based literature (Caliendo and Kopeinig 2008).

The (un-normalized) inverse probability weighting (IPW) IV-estimator can be written as

$$\widehat{\delta}_{IPW} = \frac{\sum_{i=1}^{N} \frac{Z_i Y_i}{\widehat{P}_i} - \frac{(1-Z_i)Y_i}{1-\widehat{P}_i}}{\sum_{i=1}^{N} \frac{Z_i D_i}{\widehat{P}_i} - \frac{(1-Z_i)D_i}{1-\widehat{P}_i}}, \tag{7}$$

where $\widehat{P}_i$ is an estimate of the PS $P(Z_i = 1|X_i)$. IPW has been shown to be semi-parametrically efficient in the IV context (Donald et al. 2014). To estimate the PS, two approaches are used. First, the same logit estimate as for KM is employed. To ensure better performance, weights of this estimator are normalized as un-normalized

[4] Other flexible estimators are available. See Abadie (2003), Tan (2006), MaCurdy et al. (2011) and Donald et al. (2014), Sant'Anna et al. (2022) and Sloczynski et al. (2022). The latter two are promising extensions of the so-called "kappa weighting" by Abadie (2003).

weights may yield unreliable results (Frölich 2004; Busso et al. 2014). Akin to KM, the min-max criterion is used to ensure common support. In the context of IPW, this sort of trimming may be even more important as IPW with PS close to zero or one may lead to invalid statistical inference when using the non-parametric bootstrap as is done for all estimators considered. See Heiler and Kazak (2021) for derivations and alternative bootstrap approaches or Sasaki and Ura (2022) for trimming based methods.

Second, as IPW methods may be overly sensitive to the specification of the estimated PS (Schafer and Kang 2008), this paper also uses the novel efficient covariate balancing (ECB) procedure by Heiler (2021) to estimate the PS. This approach specifies a loss-function tailored to the estimation of treatment effects using IVs and algorithmically minimizes covariate imbalances, leading to improved bias and variance properties in finite-samples compared to standard IPW methods (see Heiler, 2021, for details). This approach has several advantages. First, akin to IPW, ECB is semiparametrically efficient. Second, ECB is doubly-robust if covariates are specified flexibly and third, the ECB method tends to shrink the PS which may alleviate the need to implement heuristic trimming approaches such as the min-max criterion.[5] Due to the last property, IPW based on the ECB is implemented without further common support restrictions.

Ultimately, choosing an IV estimator involves a trade-off: Standard 2SLS is more easily applied than matching or weighting but 2SLS may be inconsistent under effect heterogeneity. Moreover, recent simulation evidence by Sloczynski et al. (2022) suggests that more flexible estimators may even be competitive in terms of mean squared error compared to standard 2SLS. However, more research on the relative performance of IV estimation methods under realistic data-generating processes is necessary to provide better guidance to researchers.

3 Re-estimating the returns to college exploiting college proximity

This Section provides empirical evidence on the relevance of potential inconsistencies in 2SLS estimates when an instrument is only valid conditional on covariates. This is done by re-estimating the wage returns to college exploiting college proximity as instrument using the data originally analyzed by Card (1995).

3.1 Data and descriptives

The data stem from the National Longitudinal Survey of Young Men, which interviewed men aged 14–24 in 1966 with follow-up surveys until 1981. The dataset contains information on 1976 log-earnings, years of education, and an indicator for growing up in a local labor market with an accredited 4-year college as well as covariates. The latter consist of potential experience, indicators for the 1966 census region, an indicator for being black, and living in the south as well as in an urban area in 1966 and 1976. Following Sloczynski (2021), a subset of the original data is analyzed with

[5] Note that ECB weights are normalized by construction if an intercept is included in the model, which is done for all analyses performed using ECB in this paper.

Table 1 Descriptive statistics

	Growing up near a 4-year college?		*p* value
	Yes $Z_i = 1$	No $Z_i = 0$	
Mean experience (years)	8.67	9.21	0.00
Share black	0.21	0.28	0.00
Share south in 1966	0.33	0.60	0.00
Share urban in 1966	0.80	0.33	0.00
Share with some college	0.55	0.42	0.00
Mean log-wages	6.31	6.16	0.00
Observations	2038	950	

p values are obtained from a *t* test of equal means

at least five observations in each covariate cell given by the interactions of the five indicators for being black, living in the south and in an urban area in 1966 and 1976. This restriction results in a sample of 2988 individuals instead of the 3010 originally analyzed by Card (1995).[6]

The main idea of the instrumental variable set-up is that children who grew up near a college may live with their parents throughout their studies and thus face lower cost of post-secondary education, which should increase the likelihood of going to college independent of their ability. Accordingly, the treatment variable is defined as "some college", i.e. having strictly more than 12 years of education.[7]

Table 1 provides some select descriptive statistics for the sample, split by whether individuals grew up near a college ($Z_i = 1$) or not ($Z_i = 0$).

The descriptive statistics reveal quite sizable differences in terms of covariate distributions between groups defined by the binary instrument. The most-striking difference can be seen in the likelihood of living in an urban area: 80% of individuals who grew up near a 4-year college lived in an urban area in 1966, whereas the same is only true for 33% among individuals who grew up without a college nearby. Similarly, individuals who lived in the south in 1966 are under-represented among individuals who grew up near a 4-year college: of those who did (not) grow up near a 4-year college, 33 (60) percent lived in the south. Moreover, differences in racial composition and experience are also non-negligible. All of these differences are highly statistically significant as indicated by the small *p* values obtained from equality of means tests. As these variables tend to show quite strong associations with the outcome of interest, it is unlikely

[6] None of the results presented are sensitive to this restriction.

[7] Note that this definition of the treatment variable is a binarized variable based on an underlying multi-valued treatment variable (i.e. years of education). Such binarization may lead to a violation of the exclusion restriction (Andresen and Huber 2021). While such violations may affect the resulting estimates, the impact of this issue should be of minor importance as this paper compares differences between estimators which are all affected by such an issue.

that the instrument is valid without conditioning on covariates and hence, an unconditional comparison of college attendance rates and log-wages across instrument groups is unlikely to be informative about the true effect of college attendance on earnings.

3.2 Specification and estimation methods

The returns to college will be estimated using two different sets of covariates. First, the main specification of Card (1995)—referred to as the baseline specification—will be used. This specification consists of potential experience in linear and squared form, indicators for the 1966 census region, an indicator for being black, and living in the south in 1976 as well as indicators for living in an urban area in 1966 and 1976. Second, following Sloczynski (2021), a saturated, i.e. fully interacted, specification based on the indicator for being black, living in the south in 1966 and 1976 and living in an urban area in 1966 and 1976 is used. Sloczynski (2021) adopted this flexible specification because Kitagawa (2015) provided evidence in favor of the validity of the instrument after conditioning on these covariates.

To estimate effects of college attendance on wages, the previously discussed methods are used. That is, naïve OLS and standard 2SLS and more flexible estimators based on the PS are applied. The latter consist of KM with an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo et al. 2008), IPW based on a logit estimate of the PS as well as ECB (Heiler 2021). When estimating effects based on the logit estimate of the PS, common support is imposed via the min–max criterion by Dehejia and Wahba (1999) as is standard in the PS-based literature (Caliendo and Kopeinig 2008).

As Sloczynski (2021) raises doubts about the validity of the strong monotonicity assumption, all estimators are also be applied using the reordered instrument. Following Sloczynski (2021), this adjusted instrument is obtained by estimating first stage effects non-parametrically for each covariate cell of the saturated specification and then reversing the instrument for individuals estimated to be defiers such that Z_i^R encourages treatment for everyone. This changes the target parameter from CACE to MACE. In order to take care of this additional estimation step when performing statistical inference, standard errors are estimated using the non-parametric bootstrap not just for the PS-based estimators but also for 2SLS when using the reordered instrument. Standard errors are obtained using 999 replications, inference is based on the normal approximation.

3.3 Implementing matching and weighting

Before turning to actual estimates, it is imperative to check overlap and common support in terms of the PS as well as covariate balancing after matching or weighting (Caliendo and Kopeinig 2008). Figure 1 shows histograms of estimated PS with a bin size of 2.5%. Visual inspection suggests sufficient overlap between instrument groups, independent of the specification and estimation procedure used. Moreover, the PS distributions appear to be sufficiently bounded away from zero or one, which is important for the non-parametric bootstrap employed to be valid (Heiler and Kazak

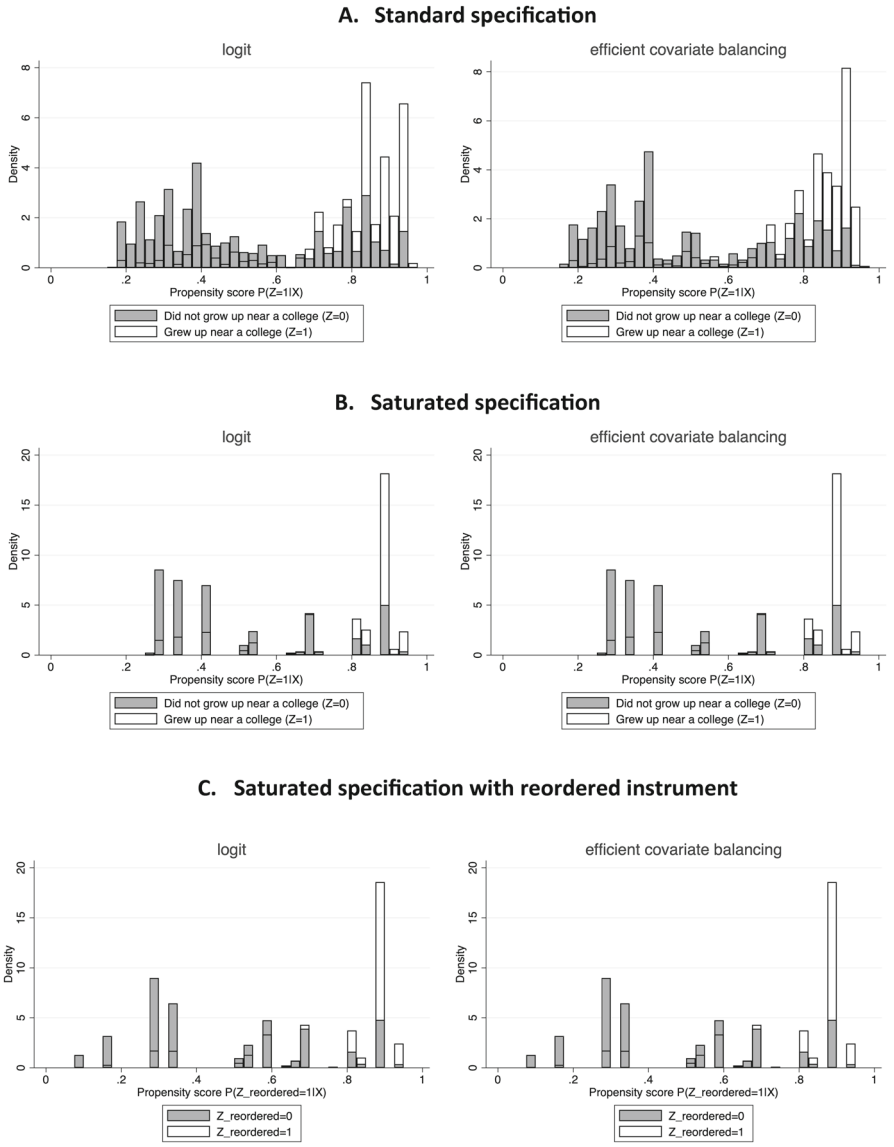

Fig. 1 Propensity score distributions. This figure shows histograms of estimated propensity scores using either a logit regression or the efficient covariate balancing procedure by Heiler (2021). The baseline specification consists of experience, experience squared and indicators for being black, living in the south, urban in 1966 and 1976 as well as census region of residence. The saturated specification consists of group dummies for the fully-interacted set of dummy variables for being black, living in the south in 1966 and 1976 as well as residence in an urban region in 1966 and 1976.

Table 2 Balancing

	Before	After balancing		
		KM	IPW	ECB
A. Baseline specification				
Pseudo-R^2	0.211	0.009	0.006	0.000
Overall p value	0.000	0.151	0.570	1.000
No. of individuals off support		36	36	
B. Saturated specification				
Pseudo-R^2	0.207	0.000	0.000	0.000
Overall p value	0.000	1.000	1.000	1.000
No. of individuals off support		0	0	
C. Saturated specification with reordered instrument				
Pseudo-R^2	0.221	0.000	0.000	0.000
Overall p value	0.000	1.000	1.000	1.000
No. of individuals off support		0	0	

Pseudo-R^2 are from a (weighted) logit regression of the (reordered) instrument on covariates. The overall p value is taken from a likelihood-ratio test on the excludability of covariates. The baseline specification consists of experience, experience squared and indicators for being black, living in the south, urban in 1966 and 1976 as well as census region of residence. The saturated specification consists of group dummies for the fully-interacted set of dummy variables for being black, living in the south in 1966 and 1976 as well as residence in an urban region in 1966 and 1976. Kernel matching (KM) is performed using an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo et al. 2008). Inverse probability weighting is based on either the logit estimate of the propensity score (IPW) or the efficient covariate balancing (ECB) score by Heiler (2021). For KM and IPW, observations off support according to the min-max criterion are discarded (Dehejia and Wahba 1999)

2021; Sasaki and Ura 2022).[8] Applying the min-max criterion for KM and IPW based on the standard specification of the logit PS leads to the exclusion of 36 individuals. This equals roughly 1.2% of the sample and thus, one should not be overly concerned that estimated effects are no longer representative of the target estimand. Regarding covariate balance, Table 2 shows the pseudo-R^2 from a logit regression before and after matching or weighting for each specification used. All balancing approaches yield a substantial reduction in imbalance from around 20% to less than 1%. As intended, ECB delivers exact balance, independent of the specification used. Moreover, p values of likelihood-ratio tests suggest that after matching or weighting, covariates are no longer statistically associated with the instrument. Hence, these statistics suggest adequate covariate balance in order to move on to the outcome analysis.

[8] The minimum and maximum PS values obtained via logit regression are 0.172 and 0.950 (baseline specification), 0.250 and 0.931 (saturated specification) and 0.086 and 0.931 (saturated specification, reordered instrument). The minima and maxima obtained via ECB are very similar to the logit estimates for the standard specification and identical for the others.

3.4 Comparing parametric and more flexible estimates of the returns to college

Focusing on the standard specification in the first two columns of Table 3, one can see that the 2SLS estimate of roughly 0.6 log-points is more than twice as large as the OLS estimate of about 0.24 log-points. This is in line with the findings of Card (1995) using multi-valued years of education as the treatment variable instead of a binary variable as in this case. Similar results are found using the saturated specification: 2SLS yields a point estimate of 0.57 log-points and the naïve OLS estimate is even smaller than when using the standard specification. Card (1995) attributes the sizable

Table 3 Main Results

	OLS	2SLS	KM	IPW	ECB
A. Baseline specification					
Return to college	0.238***	0.603**	0.279	0.323	0.289
	(0.017)	(0.289)	(0.228)	(0.219)	(0.212)
Reduced form effect		0.039**	0.028	0.033*	0.028
		(0.018)	(0.021)	(0.020)	(0.019)
First stage effect		0.065***	0.099***	0.102***	0.098***
		(0.019)	(0.024)	(0.021)	(0.020)
B. Saturated specification					
Return to college	0.111***	0.570	0.266	0.266	0.266
	(0.015)	(0.350)	(0.399)	(0.399)	(0.505)
Reduced form effect		0.034*	0.022	0.022	0.022
		(0.018)	(0.020)	(0.020)	(0.019)
First stage effect		0.059***	0.082***	0.082***	0.082***
		(0.022)	(0.023)	(0.023)	(0.023)
C. Saturated specification with reordered instrument					
Return to college	0.111***	0.289	0.192	0.192	0.192
	(0.015)	(0.195)	(0.173)	(0.173)	(0.169)
Reduced form effect		0.031	0.023	0.023	0.023
		(0.024)	(0.023)	(0.023)	(0.022)
First stage effect		0.106***	0.119***	0.119***	0.119***
		(0.017)	(0.020)	(0.020)	(0.020)

The baseline specification consists of experience, experience squared and indicators for being black, living in the south, urban in 1966 and 1976 as well as census region of residence. The saturated specification consists of group dummies for the fully-interacted set of dummy variables for being black, living in the south in 1966 and 1976 as well as residence in an urban region in 1966 and 1976. Kernel matching (KM) is performed using an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo et al. 2008). Inverse probability weighting is based on either the logit propensity score (IPW) or the efficient covariate balancing (ECB) score by Heiler (2021). Standard errors for matching and weighting estimators as well as the 2SLS estimator based on the reordered instrument are obtained via 999 bootstrap replications. Tests on statistical significance use the normal approximation. Significance at the 10/5/1% level is denoted by */**/***

gap in estimates between 2SLS and OLS to possibly higher returns to education among individuals with a relatively poor background as they are the most likely to be induced to receive additional education by the instrument. This may explain why effects are expected to be larger, but estimates appear to be unreasonably large. Sloczynski (2021) argues that the large estimate may be caused by the existence of defiers. Indeed, his results—which are replicated in Table 3—show that when accounting for the existence of defiers by using the reordered instrument, the 2SLS estimate drops substantially to around 0.29 log-points. Nonetheless, the estimated effect is still considerably larger than the effect of roughly 20% suggested by other research on the returns to college (see for example Hoekstra 2009; Smith et al. 2020; Zimmerman 2014).

Turning to the more flexible estimates based on the PS in columns three to five of Table 3, one can see that matching and weighting estimators yield substantially smaller point estimates of the returns to college than 2SLS.[9] Estimates range from 0.28 to 0.32 log-points for the baseline specification. Estimates using the saturated specification are essentially identical due to their non-parametric nature, independent of whether KM or IPW with a logit or ECB PS is used. These estimates suggest a roughly a 0.27 log-point gain in wages from college attendance. If one uses the reordered instrument instead, matching and weighting estimates drop to roughly 0.2 log-points, which is fairly close to the estimates suggested by the literature. Furthermore, Table 3 shows that these smaller point estimates of returns to college are both due to smaller reduced form estimates as well as larger first stage effects when using PS-based estimators compared to 2SLS. Overall, the results suggest that the implicit conditional-variance weighting of 2SLS may have a substantial impact on resulting effect estimates when estimating the returns to college using college proximity. 2SLS estimates are somewhere between 50 and 100% larger than more flexible PS-based estimates.[10] These differences are rather sizeable, underscoring the potential value in using more robust PS-based estimators when estimating effects using an IV set-up.

3.5 Inspecting effect heterogeneity

To further illustrate the impact of the conditional-variance weighting by 2SLS, Table 4 compares 2SLS estimates with effect estimates using PS-based estimators as well as the estimates one would obtain if one weighted PS-based estimators with an estimate of the conditional variance of the instrument, i.e. mimicking the asymptotic behavior of 2SLS. This is done for the full sample as well as for two subsamples. For the sake of brevity, results are shown only for the saturated specification with the reordered

[9] This result is also supported by contemporaneous findings by Sloczynski et al. (2022) using different versions of the kappa-approach by Abadie (2003).

[10] Note that most matching or weighting estimates of the returns to college are insignificant at common levels and that differences discussed are also not statistically different from zero due to large standard errors. The significance of differences across estimators was tested using a random sampling splitting as well as a bootstrap procedure. Both procedures lead to highly insignificant differences (results not shown, available from the author upon request). While differences may be insignificant using the sample at hand, simply focusing on statistical significance may falsely discourage the use of more robust estimation methods in favor of 2SLS in applied work.

Table 4 Effect heterogeneity—saturated specification with reordered instrument

	2SLS	KM	Variance weighted KM	IPW	Variance weighted IPW	ECB	Variance weighted ECB
A. Full sample with $P\left(Z^R = 1\right) = 0.67$							
Return to college	0.289	0.192	0.289	0.192	0.289	0.192	0.289
	(0.195)	(0.173)	(0.195)	(0.173)	(0.195)	(0.238)	(0.307)
Reduced form effect	0.031	0.023	0.031	0.023	0.031	0.023	0.031
	(0.024)	(0.023)	(0.024)	(0.023)	(0.024)	(0.031)	(0.035)
First stage effect	0.106***	0.119***	0.106***	0.119***	0.106***	0.119***	0.106***
	(0.017)	(0.020)	(0.017)	(0.020)	(0.017)	(0.023)	(0.022)
B. Urban sample with $P\left(Z^R = 1\right) = 0.79$							
Return to college	0.249	0.137	0.249	0.137	0.249	0.137	0.249
	(0.308)	(0.225)	(0.308)	(0.225)	(0.308)	(0.231)	(0.311)
Reduced form effect	0.030	0.019	0.030	0.019	0.030	0.019	0.030
	(0.041)	(0.033)	(0.041)	(0.033)	(0.041)	(0.032)	(0.039)
First stage effect	0.119***	0.136***	0.119***	0.136***	0.119***	0.136***	0.119***
	(0.026)	(0.029)	(0.026)	(0.029)	(0.026)	(0.028)	(0.025)
C. Rural sample with $P\left(Z^R = 1\right) = 0.44$							
Return to college	0.342	0.347	0.342	0.347	0.342	0.347	0.342
	(0.246)	(0.250)	(0.246)	(0.250)	(0.246)	(0.247)	(0.247)
Reduced form effect	0.031	0.031	0.031	0.031	0.031	0.031	0.031
	(0.025)	(0.026)	(0.025)	(0.026)	(0.025)	(0.025)	(0.025)

Table 4 (continued)

	2SLS	KM	Variance weighted KM	IPW	Variance weighted IPW	ECB	Variance weighted ECB
First stage effect	0.092***	0.089***	0.092***	0.089***	0.092***	0.089***	0.092***
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.024)

This table shows estimated effects for the full sample, the urban and the rural sample using the saturated specification and the reordered instrument. Living in an urban area is defined as living in a standard metropolitan area in 1966. The probability being assigned $Z^R = 1$ in the respective samples is 0.67 (full). 0.79 (urban) and 0.44 (rural). Kernel matching (KM) is performed using an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo, 2008). Inverse probability weighting is based on either the logit estimate of the propensity score (IPW) or the efficient covariate balancing (ECB) score by Heiler (2021). Variance-weighted estimators mimic the asymptotic behavior of 2SLS and weight observations by the estimated variance of the instrument obtained via logit regression. All standard errors are obtained via 999 bootstrap replications. Tests on statistical significance use the normal approximation. Significance at the 10/5/1% level is denoted by */**/***

instrument. Results for the other specifications—which are similar to the ones presented here—can be found in Tables 5 and 6 in the "Appendix".

Panel A of Table 4 first replicates estimates for 2SLS and the PS-based estimators for the full sample found in Table 3: 2SLS yields an estimate of college returns of 0.289 log-points, PS-Based estimators suggest returns of 0.192 log-points. Conditional-variance weighted KM and the other PS-based approaches yield an estimate of 0.289 log-points, which is identical to the 2SLS estimate. Thus, in the fully saturated specification, the difference between 2SLS and more flexible estimators can be entirely attributed to the conditional-variance weighting performed by 2SLS. When using a non-saturated specification, this property breaks down. However, results still clearly show that variance weighting has a major impact on resulting estimates (see Table 5).

As pointed out by one of the reviewers, results by Sloczynski (2021) imply that 2SLS is expected to yield similar estimates to more flexible estimators when the (reordered) instrument groups are roughly of the same size, i.e. when $P(Z_i = 1) \approx 0.5$ or $P(Z_i^R = 1) \approx 0.5$. To inspect this implication, Panel B and C of Table 4 estimate effects for individuals who grew up an in urban environment with $P\left(Z_i^R = 1|\text{urban}\right) = 0.79$ or in a more rural area with $P\left(Z_i^R = 1|\text{rural}\right) = 0.44$. Indeed, 2SLS estimates are much more similar to PS-based estimates in the rural sample (0.342 and 0.347 log-points) than in the urban sample (0.249 and 0.137 log-points). Again, these differences are completely accounted for by the conditional-variance weighting. Hence, it appears that 2SLS is expected to yield estimates close to more flexible estimators when instrument groups are roughly equal size *because* the conditional-variance weighting plays less of a role in that case.

4 Conclusion

By re-examining the Card (1995) data on college proximity and the returns to college, this paper shows that potential inconsistencies in 2SLS estimates of local treatment effects documented in the theoretical literature are not merely a hypothetical threat when effects are heterogenous. For the data at hand, 2SLS yields systematically larger effects than more flexible estimators based on the PS with differences amounting to roughly 50 to 100%. It is shown that this is because standard linear-in-covariates 2SLS yields a conditional variance-weighted average effect, putting more weight on units with a PS close to a coin flip. In line with theoretical predictions by Sloczynski (2021), the results suggest that 2SLS estimates can be expected to be more trustworthy when sample shares of instrument groups are roughly of equal size. Moreover, the paper shows that this is because the effects of conditional-variance weighting tend to be less severe when groups sizes are similar. Overall, the results show that the presumption that 2SLS yields point estimates close to more flexible estimators based on the PS as argued by Angrist and Pischke (2008) does not apply in general and that one should be suspicious of 2SLS estimates when group sizes differ substantially and covariates are predictive of the instrument. In that case it may be best to use semi- or non-parametric estimation techniques instead. At the very least, one should use these methods to assess the sensitivity of estimates regarding implicit parametric assumptions made when using linear-in-covariates 2SLS.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-023-02441-7.

Acknowledgements The author would like to thank the reviewers for the very helpful comments which lead to a substantial improvement of the paper during the revision.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or. animals performed by any of the authors.

Appendix

See Tables 5 and 6.

Table 5 Effect heterogeneity – baseline specification

	2SLS	KM	Variance weighted KM	IPW	Variance weighted IPW	ECB	Variance weighted ECB
A. Full sample with P(Z = 1) = 0.68							
Return to college	0.603**	0.182	0.314	0.323	0.590	0.289	0.559
	(0.289)	(0.243)	(0.548)	(0.219)	(0.505)	(0.212)	(9.430)
Reduced form effect	0.039**	0.021	0.023	0.033*	0.037**	0.028	0.032*
	(0.018)	0.024	(0.021)	(0.020)	(0.018)	(0.019)	(0.019)
First stage effect	0.065***	0.117***	0.073***	0.102***	0.063***	0.098***	0.058***
	(0.019)	(0.025)	(0.023)	(0.021)	(0.018)	(0.020)	(0.019)
B. Urban sample with P(Z = 1) = 0.84							
Return to college	0.376	0.092	0.422	0.027	0.394	0.172	0.349
	(0.282)	(0.199)	(0.701)	(0.177)	(0.513)	(0.218)	(0.339)
Reduced form effect	0.034	0.014	0.037	0.005	0.035	0.022	0.031
	(0.026)	(0.031)	(0.028)	(0.029)	(0.026)	(0.027)	(0.025)
First stage effect	0.091***	0.152***	0.088***	0.172***	0.089***	0.129***	0.089***
	(0.027)	(0.031)	(0.030)	(0.028)	(0.027)	(0.026)	(0.025)
C. Rural sample with $P(Z = 1) = 0.39$							
Return to college	1.032	0.867	1.024	0.969	1.129	1.198	1.031
	(0.607)	(41.18)	(7.288)	(17.16)	(26.79)	(7.913)	(6.251)

Table 5 (continued)

	2SLS	KM	Variance weighted KM	IPW	Variance weighted IPW	ECB	Variance weighted ECB
Reduced form effect	0.055**	0.054*	0.055*	0.049*	0.049**	0.064**	0.055*
	(0.026)	(0.030)	(0.029)	(0.027)	(0.025)	(0.027)	(0.026)
First stage effect	0.054***	0.062*	0.053*	0.050*	0.043	0.054*	0.053**
	(0.026)	(0.033)	(0.031)	(0.030)	(0.027)	(0.028)	(0.027)

This table shows estimated effects for the full sample, the urban and the rural sample using the baseline specification and the original instrument. Living in an urban area is defined as living in a standard metropolitan area in 1966. The probability of growing up near a four-year college $P(Z = 1)$in the respective samples is 0.68 (full), 0.84 (urban) and 0.39 (rural). Kernel matching (KM) is performed using an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo, 2008). Inverse probability weighting is based on either the logit estimate of the propensity score (IPW) or the efficient covariate balancing (ECB) score by Heiler (2021). Variance-weighted estimators mimic the asymptotic behavior of 2SLS and weight observations by the estimated variance of the instrument obtained via logit regression. Standard errors of matching and weighting estimates are obtained via 999 bootstrap replications. Tests on statistical significance use the normal approximation. Significance at the 10/5/1% level is denoted by */**/***. ECB estimates for sub-samples are obtained using covariate-adjusted regressions on the re-weighted samples due to insufficient balance after weighting

Table 6 Effect heterogeneity—saturated specification

	2SLS	KM	Variance weighted KM	IPW	Variance-weighted IPW	ECB	Variance-weighted ECB
A. Full sample with P(Z = 1) = 0.68							
Return to college	0.570	0.266	0.570	0.266	0.570	0.266	0.570
	(0.350)	(0.399)	(1.419)	(0.399)	(1.419)	(0.425)	(5.136)
Reduced form effect	0.034*	0.022	0.034*	0.022*	0.034*	0.022*	0.034*
	(0.018)	(0.020)	(0.018)	(0.020)	(0.018)	(0.020)	(0.018)
First stage effect	0.059***	0.082***	0.059***	0.082***	0.059***	0.082***	0.059***
	(0.022)	(0.023)	(0.021)	(0.023)	(0.021)	(0.024)	(0.021)
B. Urban sample with P(Z = 1) = 0.84							
Return to college	0.350	0.116	0.350	0.116	0.350	0.116	0.350
	(0.335)	(0.323)	(1.538)	(0.323)	(1.538)	(0.344)	(1.578)
Reduced form effect	0.029	0.012	0.029	0.012	0.029	0.012	0.029
	(0.027)	(0.027)	(0.026)	(0.027)	(0.026)	(0.026)	(0.025)
First stage effect	0.082***	0.107***	0.082***	0.107***	0.082***	0.107***	0.082***
	(0.031)	(0.032)	(0.031)	(0.032)	(0.031)	(0.031)	(0.030)
C. Rural sample with P(Z = 1) = 0.39							
Return to college	1.099	1.107	1.099	1.107	1.099	1.107	1.099
	(1.054)	(245.2)	(11.49)	(245.4)	(11.49)	(27.40)	(17.25)
Reduced form effect	0.039	0.039	0.039	0.039	0.039	0.039	0.039
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)

Table 6 (continued)

	2SLS	KM	Variance weighted KM	IPW	Variance-weighted IPW	ECB	Variance-weighted ECB
First stage effect	0.035	0.036	0.035	0.036	0.035	0.036	0.035
	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.029)	(0.029)

This table shows estimated effects for the full sample, the urban and the rural sample using the saturated specification and the original instrument. Living in an urban area is defined as living in a standard metropolitan area in 1966. The probability of growing up near a four-year college P(Z = 1) in the respective samples is 0.68 (full), 0.84 (urban) and 0.39 (rural). Kernel matching (KM) is performed using an Epanechnikov kernel using a bandwidth chosen via weighted cross-validation (Galdo, 2008). Inverse probability weighting is based on either the logit estimate of the propensity score (IPW) or the efficient covariate balancing (ECB) score by Heiler (2021). Variance-weighted estimators mimic the asymptotic behavior of 2SLS and weight observations by the estimated variance of the instrument obtained via logit regression. Standard errors of matching and weighting estimates are obtained via 999 bootstrap replications. Tests on statistical significance use the normal approximation. Significance at the 10/5/1% level is denoted by */**/***

References

Abadie A (2003) Semiparametric instrumental variable estimation of treatment response models. J Econ 113(2):231–263

Abadie A, Cattaneo MD (2018) Econometric methods for program evaluation. Annual Rev Econ 10:465–503

Andresen M, Huber M (2021) Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction. Econom J 24(3):536–558

Angrist JD (1998) Estimating the labor market impact of voluntary military service using social security data on military applicants. Econometrica 66(2):249–288

Angrist JD, Pischke JS (2008) Mostly harmless econometrics. Princeton UUniversity Press, Princeton

Austin PC, Stuart EA (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 34(28):3661–3679

Blackburn ML, Neumark D (1993) Omitted-ability bias and the increase in the return to schooling. J Labor Econ 11(3):521–544

Blandhol C, Bonney J, Mogstad M, Torgovitsky A (2022) When is TSLS actually late? NBER working paper no. 29709

Busso M, DiNardo J, McCrary J (2014) New evidence on the finite sample properties of propensity score reweighting and matching estimators. Rev Econ Stat 96(5):885–897

Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. J Econ Surv 22(1):31–72

Card D (1995) Using geographic variation in college proximity to estimate the return to schooling. In: Christophides LN, Grant EK, Swidinsky R (eds) Aspects of labour market behavior: essays in Honour of John Vanderkamp. University of Toronto Press, Toronto, pp 201–222

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94(448):1053–1062

Donald SG, Hsu Y, Lieli RP (2014) Inverse probability weighted estimation of local average treatment effects: a higher order MSE expansion. Stat Probab Lett 95:132–138

Frölich M (2004) Finite-sample properties of propensity-score matching and weighting estimators. Rev Econ Stat 86(1):77–90

Frölich M (2007) Nonparametric IV estimation of local average treatment effects with covariates. J Econ 139(1):35–75

Galdo JC, Smith J, Black D (2008) Bandwidth selection and the estimation of treatment effects with unbalanced data. Ann d'Économie et de Statistique 91/92:189–216

Heiler P, Kazak E (2021) Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. J Econ 222(2):1083–1108

Heiler P (2021) Efficient covariate balancing for the local average treatment effect. J Bus Econ Stat (**forthcoming**)

Hoekstra M (2009) The Effect of attending the Flagship State University on earnings: a discontinuity-based approach. Rev Econ Stat 91:717–724

Huber M, Mellace G (2015) Testing instrument validity for LATE identification based on inequality moment constraints. Rev Econ Stat 97(2):398–411

Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–476

Kitagawa T (2015) A test for instrument validity. Econometrica 83(5):2043–2063

Kolésar M (2013) Estimation in an instrumental variables model with treatment effect heterogeneity. Unpublished manuscript

MaCurdy T, Chen X, Hong H (2011) Flexible estimation of treatment effect parameters. Am Econ Rev 101(3):544–551

Mourifié I, Wan Y (2017) Testing local average treatment effect assumptions. Rev Econ Stat 99(2):305–313

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688

Sant'Anna PH, Song X, Xu Q (2022) Covariate distribution balance via propensity scores. J Appl Econ 37(6):1093–1120

Sasaki Y, Ura T (2022) Estimation and inference for moments of ratios with robustness against large trimming bias. Econ Theory 38(1):66–112

Schafer JL, Kang J (2008) Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods 13(4):279
Sloczynski T (2021) When should we (not) interpret linear IV estimands as LATE? IZA discussion papers 14349, Institute of Labor Economics (IZA)
Sloczynski T, Uysal SD, Wooldridge JM (2022) Abadie's kappa and weighting estimators of the local average treatment effect. CESifo working paper no. 9715
Smith J, Goodman J, Hurwitz M (2020) The economic impact of acccess to public four-year colleges, NBER Working Paper no. 27177
Tan Z (2006) Regression and weighting methods for causal inference using Instrumental variables. J Am Stat Assoc 101(476):1607–1618
Thoemmes FJ, Kim ES (2011) A systematic review of propensity score methods in the social sciences. Multivar Behav Res 46(1):90–118
Wooldridge JM (2010) Econometric analysis of cross section and panel data, vol 1, 2nd ednMIT Press Books, The MIT Press, Cambridge
Zimmerman SD (2014) The returns to College Admission for academically marginal students. J Labor Econ 32:711–754