Instrumental Variables and Causal Mechanisms: Unpacking the Effect of Trade on Workers and Voters*

Christian Dippel[†] Robert Gold[‡] Stephan Heblich[§] Rodrigo Pinto[¶]

August 25, 2017

Abstract

In this paper we pursue an instrumental variable (IV) strategy to study the effect of trade exposure on voters' support for populist parties, using German data from 1987–2009. We find that support for extreme right parties increased in regions that faced more import competition. These regions also experienced significant labor market turmoil. The focus of this paper is on the extent to which trade exposure polarized voters *because* it caused labor market turmoil. To answer this question we propose a new framework for mediation analysis in IV settings. Our main finding is that the effect of import exposure on labor markets *entirely* explains its polarizing effect on voters. Our empirical framework may be useful in a broad range of empirical applications studying causal mechanisms in IV settings.

Keywords: Instrumental Variables, Trade Exposure, Voting, Local Labor Markets, Causal Mediation Analysis

JEL Codes: D72, F6, J2

[*] We thank Alberto Alesina, Sascha Becker, Gilles Duranton, Jon Eguia, Andreas Ferrara, Paola Giuliano, Tarek Hassan, Kosuke Imai, Ed Leamer, Yi Lu, Ralph Ossa, Anne Otto, David Pacini, Nicola Persico, Maria Petrova, Justin Pierce, Giacomo Ponzetto, Daniel Sturm, Peter Schott, Zhigang Tao, Dan Trefler, Dan Treisman, Nico Voigtländer, Wouter Vermeulen, Till von Wachter, Romain Wacziarg, Frank Windmeijer, Yanos Zylberberg, and seminar participants at Bristol, Kiel, the LMU, the LSE, Toronto, UCLA, U Hong Kong, Warwick, the 2013 Urban Economics Association and German Economists Abroad Meetings, the 2015 Quebec Political Economy conference, and the 2016 EEA conference for valuable comments and discussions. We also thank David Slichter for thoughtful comments and Wolfgang Dauth for sharing the crosswalk from product classifications to industry classifications in the German IAB data. Dippel acknowledges financial support from UCLA's Center for Global Management. This paper builds on an earlier working paper titled "Globalization and its (Dis-)Content."

[†] University of California, Los Angeles, CCPR, and NBER.

[‡] IfW - Kiel Institute for the World Economy.

[§] University of Bristol, CESifo, IZA, and SERC.

[¶] University of California, Los Angeles, CCPR, and NBER.

1 Introduction

International trade between high and low-wage countries has risen dramatically in the last thirty years (Krugman, 2008). While consumers in high-wage countries have benefited from cheaper manufacture goods, there has also been real wage stagnation and substantial losses of manufacturing jobs (Autor, Dorn, and Hanson, 2013; Dauth, Findeisen, and Suedekum, 2014; Pierce and Schott, 2016; Malgouyres, 2017). This dichotomy has fostered political polarization and the rise of parties and politicians with protectionist agendas (Malgouyres, 2014; Feigenbaum and Hall, 2015; Autor, Dorn, Hanson, and Majlesi, 2016; Jensen, Quinn, and Weymouth, 2016; Che, Lu, Pierce, Schott, and Tao, 2016).

We identify and evaluate the causal mechanism that links the nexus of international trade, labor markets and political change. To know the extent to which trade exposure (T) polarized voters *because* it caused labor market turmoil, one first needs to estimate the causal effect of T on labor markets (M) and on voting (Y). The identification challenge is that T is endogenous because of *unobserved confounding variables* V that cause both T and M and Y. For example, a common concern in the literature above (where local trade exposure is measured as often a function of local industry structure) is that domestic industry-level demand conditions determine local trade exposure but also influence labor markets and voter behavior. We follow the work of Autor et al. (2013) who suggest the trade exposure of high-wage countries other than Germany as an IV for Germany's trade exposure T. The resulting identifying variation is driven by supply changes (productivity or market access increases) in low-wage countries instead of fluctuations in German domestic conditions. For our purposes, we take the validity of this instrument as given and refer the reader to the literature above for a detailed discussion of its appeal.

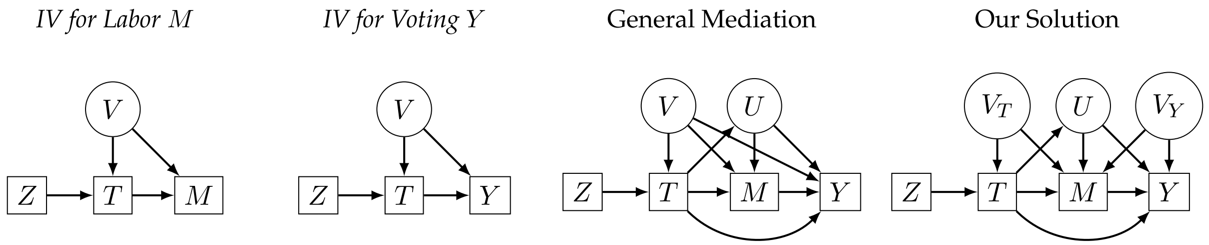
The IV generates two exclusion restrictions: $M(t) \perp\!\!\!\perp Z$ and $Y(t) \perp\!\!\!\perp Z$.[1] These two restrictions enable the identification of the causal effects of T on M and of T on Y. No distinction is made in the assumptions needed to identify the two effects. This is common and we note in passing that three pairs of papers in the literature above each use the same identification strategy to separately investigate the effect of T on labor markets and on some form of political outcomes.[2]

[1] $Y(t)$ stands for the 'counterfactual outcome' and $M(t)$ for the 'counterfactual mediator' when T is fixed at value t. See section 4 for detailed notation.

[2] e.g. Autor et al. (2013) and Autor et al. (2016), Malgouyres (2017) and Malgouyres (2014), as well as Pierce and Schott (2016) and Che et al. (2016).

Table 1: General Mediation Model and Our Solution to the Identification Problem

A. Directed Acyclic Graph (DAG) Representation



B. Structural Equations

		$T = f_T(V, Z, \epsilon_T)$	$T = f_T(Z, V_T, \epsilon_T)$
$T = f_T(Z, V, \epsilon_T)$,	$T = f_T(Z, V_T, \epsilon_T)$,	$U = f_U(T, \epsilon_U)$	$U = f_U(T, \epsilon_U)$
$M = f_M(T, Z, \epsilon_M)$,	$Y = f_Y(T, Z, \epsilon_Y)$,	$M = f_M(T, U, V, \epsilon_M)$	$M = f_M(T, U, V_T, \epsilon_M)$
Z,V, ϵ_T, ϵ_M Stat. Indep.	Z,V, ϵ_T, ϵ_Y Stat. Indep.	$Y = f_Y(T, M, U, V, \epsilon_Y)$	$Y = f_Y(T, M, U, V_Y, \epsilon_Y)$
		Z,$V, U, \epsilon's$ Stat. Indep.	Z,$V_Y, V_T, \epsilon's$ Stat. Indep.

The first and second columns of Panel A give the directed acyclic graph (DAG) representation of the two IV Models, which enable the identification of the causal effects of T on M and T on Y. The third column presents the General Mediation Model with an instrumental variable Z. Identification can only be achieved with an additional designated instrument for M (not depicted). The fourth column of Panel A presents the Restricted Mediation Model with an instrumental variable Z. This model enables the identification of the total, the direct and the indirect effect of T on Y. The model also enables the identification of the causal effect of M on Y. Panel B presents the nonparametric structural equations of each model. Conditioning variables are suppressed for sake of notational simplicity. We refer to Heckman and Pinto (2015a) for a recent discussion on causality and directed acyclic graphs.

The first and second column of Table 1 describe the two IV models. Without additional assumptions, the extent to which T causes Y *through* M remains unidentified in these two models. Our question requires a *mediation analysis*, whereby trade T causes an intermediate outcome M (labor markets) which in turn causes a final outcome Y (voting).[3] Mediation analysis decomposes the *total effect* of T on Y into the effect of T on Y that operates through M – the *indirect effect* – and the residual effect that does not – *the direct effect*. The third column of Table 1 shows the identification challenges in combining the two IV models into a *General Mediation Model*. The relation among T, M, Y may be tempered by confounding variables V that jointly cause T, M, Y. In addition, it may be tempered by unobserved mediators U that are caused by T and cause not only Y but also M.[4]

[3] For recent works on this literature, see (Heckman and Pinto, 2015b; Pearl, 2014; Imai, Keele, and Tingley, 2010).

[4] A pertinent example (in the U.S. political context) are *Trade Adjustment Assistance* (TAA) programs, which are specifically designed to cushion the effect of trade exposure on labor markets and which may well impact the vote share of the Democratic party that has championed these programs. Similar programs exist in Germany.

Two approaches exist in the literature to gain identification in this model. One approach is to focus on randomized control trials where T is randomly assigned (i.e. there is no V and no Z), and to additionally assume that there are no unobserved mediators U.[5] A second approach requires having additional dedicated instruments for M (Frölich and Huber, 2014).

Our approach differs from both in that we are explicitly interested in studying an endogenous T but want to focus on an IV setting in which we have dedicated instruments only for T. We view this as a the most common research setting with observational data.

To achieve identification, we replace the confounding variable V that jointly causes T, M, Y by two unobserved variables: V_T that causes T and M and V_Y that causes M and Y. This assumption partitions the confounding effects of V into the confounding effect of V_T that affects the correlation between T and M and V_Y that affects M and Y. This causal assumption does not eliminate the problem of confounding effects; indeed variables T, M, Y remain endogenous. Yet, section 4 shows that this assumption generates the exclusion restriction $Y(t) \perp\!\!\!\perp Z|T$ that allows us to identify the causal effect of M on Y. The resulting model is presented in the last column of Table 1.

Effectively, we assume that any confounders that influence T and Y in fact influence Y only though M. In our empirical context, if we think that domestic industry-level demand conditions are important confounders, then the framework assumes that these impact voters only to the extent that they manifest in local labor markets, i.e. as V_T. At the same time, the framework allows for local shifters in the political climate V_Y that influence voting and local labor markets but not a region's exposure to trade. Given the space constraints of an introduction, we refer the reader to section 4.2 for further discussion of the framework's intuition, and explain why our causal assumption is plausible in our empirical context. We also discuss research questions for which our identifying assumption seems less plausible. Plausibility aside, we also show that our causal assumption is testable and provide a simple model specification test based on a comparison of the general and restricted versions of the mediation model.

Lastly, under linearity, we show that our model can be estimated by standard Two-Stage Least Squares (TSLS). We ensure that we pass the model specification test and estimate the causal effects

[5] A large literature evaluates causal effects in mediation models without the use of instrumental variables. For examples of this literature, see Imai, Keele, and Yamamoto (2010); Imai, Keele, Tingley, and Yamamoto (2011a,b). This literature advances the exogeneity assumptions invoked by matching methods (Imbens, 2004; Heckman, 2008).

of T on M (i.e. $\hat{\Gamma}_T^M$), T on Y (i.e. $\hat{\Gamma}_T^Y$), and M on Y (i.e. $\hat{\Gamma}_M^{Y|T}$) using three straightforward TSLS estimations. Our main focus is the comparison of the *indirect effect* $\hat{\Gamma}_T^M \times \hat{\Gamma}_M^{Y|T}$ and the *total effect* $\hat{\Gamma}_T^Y$.

Our data combine changes in German sector-specific trade flows with local labor markets' (*Landkreis*') initial industry mix to determine regional trade exposure (T). We then instrument this with a measure based on other high-wage countries' sector-specific trade flows (Z). The data is organized as a stacked panel of two first differences for the periods 1987–1998 and 1998–2009. Each of the two periods includes a large exogenous shock to the global trading environment: In 1989, the fall of the Iron Curtain opened up the Eastern European markets, and in 2001 China's accession to the WTO led to another large increase in trade exposure. The data precede the onset of the European financial crisis. The specific start- and endpoints in our panel are dictated by national election dates. As a starting point to assess voting behavior Y we consider the entire political spectrum, before focusing on the extreme right.[6] We also use Germany's Socioeconomic Panel (SOEP) to corroborate our main *Landkreis*-level results with individual worker-level data. The SOEP is unique among similar attitudinal socio-economic surveys in having had a panel structure since the 1980s, and in that it surveys party preferences as well as workers' industry and educational background. This allows us to mirror the set-up of our main analysis, relating decadal changes in workers' stated party preferences to changes in their *Landkreis*' trade exposure over the same time.

Using regional data, we find that trade exposure (T) increased voters' support for only the narrow segment of the highly populist extreme right, with no significant effect on turnout, any of the mainstream parties, small parties, or the far left. These findings are closely corroborated by the SOEP's individual-level data, where we can also decompose the main finding by worker's type. We find that the effects are entirely driven by low-skill workers employed in manufacturing, i.e. those most affected by labor market disturbances caused by increasing international trade. We also re-affirm previous studies' findings that trade exposure (T) caused significant labor market disturbances (M): import competition reduces manufacturing's employment share, manufacturing wages, and total employment, as well as increasing unemployment. Results based on gravity

[6] Election outcomes are divided into changes in the vote-share of (i) four mainstream parties: the CDU, the SPD, the FDP and the Green party, (ii) extreme-right parties, (iii) far-left parties, (iv) other small parties, and (v) turnout, see Falck, Gold, and Heblich (2014).

residuals instead of IV are similar.[7], [8]

The empirical findings suggest that trade exposure caused changes in voting behavior *because* it affected labor markets. While this is plausible and even likely, we cannot ascertain how important a mechanism labor markets actually were because there are also other potentially important channels linking trade exposure to politics: For example, output price reductions, increases in product variety, or merely consuming more internationally-made products may all have been politically moderating. So too, may be government programs targeting trade-exposed regions. If these forces are important, then import competition may cause both political polarization and moderation. On the other hand, there is evidence that even in the absence of tangible labor market changes import competition may cause anxiety and insecurity that polarizes voters (Mughan and Lacy, 2002; Mughan, Bean, and McAllister, 2003). If these forces are important, then the populist response to trade exposure may be the result of several polarizing mechanisms, labor market adjustments being just one among them.

It is therefore necessary to estimate the extent to which labor market adjustments drive the political response to trade exposure, and we apply our framework to this. To stick to a single mediator, we focus on the *principal component* of a range of observed labor market outcomes in our data.[9] Having passed the specification test of our model's assumption, we find that the ('indirect') effect of trade exposure that is mediated by labor markets entirely explains the total effect on extreme right party votes, i.e. labor markets adjustments are clearly the most important reason for the political backlash against free trade. In fact, the *indirect* effect is somewhat larger than the *total*, implying that other effects of trade exposure were in combination politically somewhat moderating.

We contribute to the literature on causal mechanisms by offering a general mediation model that allows for endogenous variables caused by confounders and unobserved mediators. An important related paper in this literature is Frölich and Huber (2014) who investigate mediation

[7] We report these for completeness in the appendix because this is common practice but our focus is on the IV setting to which our identification framework applies.

[8] We also find that both labor markets and voting were consistently more responsive to trade shocks in the second period (1998–2009) when Germany was less regulated (Dustmann, Fitzenberger, Schönberg, and Spitz-Oener, 2014).

[9] Our framework can incorporate additional mediators, but these would require additional instruments, each designated specifically to one mediator. Because it is unlikely in observational data to have equally attractive instruments for a number of endogenous variables, our focus is on settings with designated instruments for only T, and therefore a single mediator.

models when there are separate dedicated instruments for T and M. By contrast, our model relies on an a single instrumental variable Z that directly causes T to identify three causal effects. This parsimonious feature is particularly useful for empirical applications in which good instrumental variables are scarce. Our framework can therefore be useful in a broad range of empirical applications studying causal mechanisms in IV settings. Additional advantages of our model are that it is testable and can be estimated using well-understood Two-stage Least Squares methods.

Substantively, our paper relates to three strands of the political economy literature. First, we relate to the recent body of literature on the effect of trade exposure on politics and on labor markets. Our mediation analysis framework provides a way of connecting these papers' results.[10] Second, our paper speaks to a broader literature on the effects of economic shocks on a range of political outcomes including an incumbent's reelection chances (Bagues and Esteve-Volart 2014, Jensen et al. 2016), turnout (Charles and Stephens, 2013), and stated voter preferences for redistribution (Brunner, Ross, and Washington 2011, Giuliano and Spilimbergo 2014). Third, our paper relates to an earlier political science literature on trade and political cleavages (Rogowski, 1987). This literature focuses on political cleavages along factor (e.g. occupation) or industry lines and either studies self-reported party preferences amongst voters in survey data (Scheve and Slaughter, 2001) or legislators' voting records on certain types of legislation (Hiscox, 2002). See Rodrik (1995) for an extensive survey.

In the following, section 2 describes the data. Section 3 presents the IV results establishing the causal effects of import competition on labor markets and on voting, including a micro-level analysis of workers in the SOEP. Section 4 explains our mediation model and lays out our identification approach to unpack the causal mechanism by which import competition changes voting. We refer the reader to section 4.2 for an informal discussion of the model's intuition. Section 5 applies our mediation model to unpacking the causal links between trade exposure, labor market disturbances and voting behavior. Section 6 concludes.

[10] Dix-Carneiro, Soares, and Ulyssea (2017) follow an alternative approach that is similar in spirit: they investigate the separate time-paths that labor market conditions and crime took in Brazil after trade liberalization to argue that the former response to trade caused the latter.

2 Data Construction

It is well documented that the rise of China and Eastern Europe had pronounced effects on manufacturing employment in high-wage countries. This paper studies the political consequences of this development. In Germany imports from *and* exports to China and Eastern Europe roughly tripled over the period 1987 to 1998 (from about 20 billion to about 60 billion Euros each),[11] and again tripled between 1998 and 2009. At the same time, we observe an increase in political polarization and in labor market disturbances. With a view towards the mediation framework we develop in section 4, we need the following variables to test the extent to which trade-induced labor market disturbances caused changes in voting behavior: *Treatment* T_{it} is our measure of a local labor market's trade exposure. *Mediators* M_{it} are labor market variables, and *Final Outcome* Y_{it} refers to voting outcomes. Finally, we use other countries' trade exposure to construct Z_{it} as an *Instrument* for T_{it}. We now explain how we measure these variables.

2.1 Trade Exposure (Treatment T)

To measure changes in a local labor market i's trade exposure T at time t, we follow Autor et al. (2013) and Dauth et al. (2014) and calculate

$$T_{it} = \sum_{j} \frac{L_{ijt}}{L_{jt}} \frac{\Delta IM_{Gjt} - \Delta EX_{Gjt}}{L_{it}}. \tag{1}$$

The intuition is straightforward: local labor market i's composition of employment across 157 manufacturing industries j at the beginning of period t determines its exposure to changes in industry-specific trade flows over the ensuing decade. Sector j receives more weight if region i's national share of that sector $\frac{L_{ijt}}{L_{jt}}$ is high, but a lower weight if i's overall workforce L_{it} is larger. Autor et al. (2013) focus on imports (ΔIM_{Gjt}) and consider the net of imports (ΔIM_{Gjt}) minus exports (ΔEX_{Gjt}) only in their appendix. However, Dauth et al. (2014) show that in Germany's case it is more appropriate to focus on the net, because Germany' trade with the low-wage manufacturing countries is much more balanced than the U.S.'s.

Our units of observation are counties (*Landkreise*) as a representation of local labor markets and

[11] Throughout the paper, we report values in thousands of constant-2005 Euros using exchange rates from the German *Bundesbank*.

we consider decadal changes $\Delta \mathit{Trade}_{Gjt}$. We link sectoral employment data L_{ij}, which the *Institut für Arbeitsmarkt- und Berufsforschung* (IAB) reports in standard international trade classification (SITC), to the UN Comtrade trade data using the crosswalk described in Dauth et al. (2014).

One concern with the measure of trade exposure in (1) is that it is a composite effect of the relative importance of trade-intensive industries *and* the relative importance of manufacturing employment in a region (e.g. $\frac{1}{L_{it}} \sum_j L_{ijt}$). The former determines region i's exposure to trade with low-wage countries while the latter might independently affect labor-market and voting outcomes. This problem is well known, and is solved by always conditioning on region i's initial share of manufacturing employment in all our regressions; see Autor et al. (2013).

2.2 Labor Market Variables (Mediator M)

Our labor market data stem from the *Institut für Arbeitsmarkt- und Berufsforschung* (IAB)'s Historic Employment and Establishment Statistics (HES) database (see Bender, Haas, and Klose 2000 for a detailed description). The HES is collected for social insurance purposes, and includes information on daily wages, a range of socio-demographic variables (such as educational attainment, gender, and age) and the industry, occupation, and place of work for all German workers subject to social insurance.[12] From the individual-level data we aggregate up to the *Landkreis* level to match our voting data. We consider decadal changes in six *Landkreis*-level labor market variables: total employment, manufacturing industries' share of total employment, manufacturing and non-manufacturing wages, unemployment, and finally total population size (with the latter two being provided by the German Statistical office). To summarize them more concisely, we calculate the principal components of these labor market variables and use these as our measures of labor market disturbances M_{it} in our mediation analysis. Online Appendix A provides additional information on data sources and variable construction.

[12] Civil servants and self-employed individuals are not included in our database. Furthermore, we choose to exclude workers younger than 18 or older than 65 and we exclude all individuals in training and in part-time jobs because their hourly wages cannot be assessed.

2.3 Voting (Final Outcome Y)

To measure how trade integration affects voting behavior, we focus on party-votes in federal elections in Germany (*Bundestagswahlen*).[13] Due to its at-large voting system Germany, like most continental European countries, has consistently had a multi-party system that spans the full spectrum from far-left to extreme-right parties. There are four parties that we label 'established' in that they were persistently represented in parliament over the 25 years we study. There is also a large number of small parties that run for election. The average vote share of these small parties is far below the 5% threshold of party votes needed to enter the federal parliament.[14] We collected these data to create a novel dataset of party vote shares at the county level. We group the small parties into three categories: far-left parties, extreme-right parties, and a residual category of other small parties. Altogether, we consider eight *Landkreis*-level voting outcomes as the set of final outcomes Y_{it}: changes in the vote-share of each of the four mainstream incumbent parties; changes in the aggregate vote shares of each of the far-left, extreme-right and other small parties; and finally, changes in voter turnout. Online Appendix B provides additional information on the data sources and more details on the variable construction.

2.4 The Instruments Z

There are two potential sources of bias which our measure of trade exposure (T_{it}) suggested in (1) may be vulnerable to, even once we condition on overall manufacturing employment in a region.

First, and most important, industry-specific changes in trade flows might be the result of unobserved industry demand or supply shocks in Germany, which simultaneously affect trade exposure, local labor market conditions, and as a consequence local voting behavior. For example, negative shocks to German product demand could both reduce imports (T_{it}) and demand for manufacturing workers in the affected sectors. As a consequence, OLS would underestimate the negative consequences on local labor markets in Germany and this bias could carry on to voting

[13] The party vote, called (*Zweitstimme*), mainly determines a party's share of parliamentary seats. German voters also cast a second vote for individual candidates, called (*Erststimme*). This vote for individuals affects the very composition of party factions in the parliament, but has no significant influence on their overall parliamentary share. Moreover, the decision on individual candidates might be strategic. We thus follow Falck et al. (2014) and focus on the party vote.

[14] This threshold is not binding if a party wins at least three seats through the vote for individual candidates (*Erststimme*). During our period of analysis, this occurred once in 1994. The individual candidates of the party PDS won 4 seats by *Erststimme*. As a result, the party received 30 seats in total, according its 4.4% of party votes (*Zweitstimme*) received.

behaviour.

To overcome this problem, we follow the approach in Autor et al. (2013) and instrument Germany's imports from (exports to) China or Eastern Europe, ΔIM_{Gjt} (ΔEX_{Gjt}), with the average imports from (exports to) a set of similar high-wage economies 'O', ΔIM_{Ojt} (ΔEX_{Ojt}).[15] A second concern is that reverse causality may bias our estimations if the anticipation of future import competition or export opportunities affected contemporaneous employment. To account for this, we lag the initial employment share in sectors j and regions i and the initial workforce by one decade and denote this lag by the subscript $t-1$. Combining the second and third argument, we derive the instruments

$$Z_{it}^{IM} = \sum_j \frac{L_{ijt-1}}{L_{jt-1}} \frac{\Delta IM_{Ojt}}{L_{it-1}} , \qquad Z_{it}^{EX} = \sum_j \frac{L_{ijt-1}}{L_{jt-1}} \frac{\Delta EX_{Ojt}}{L_{it-1}} . \tag{2}$$

We define separate import and export instruments instead of combining them because having two designated instruments for T enables us to perform the specification test for our model; see Section 4.4.

2.5 Space- and Time-Dimension

All the data for our main empirical analyses is observed at the county (*Landkreis*) level. We observe 408 counties in our data, 86 of which are in East Germany. Indeed, it is a unique feature of the German data that it allows assessing trade exposure (T_{it}), local labor market disturbances (M_{it}), and voting behavior (Y_{it}) in the same spatial unit. We organize our data as stacked panel of first differences between election dates. Accordingly, we deviate slightly from studying decennial changes, and instead study two periods of 11 years, 1987 to 1998 (period 1) and 1998 to 2009 (period 2). Our analysis starts with the federal election of 1987, i.e. before the fall of the Iron Curtain in 1989 and Germany's subsequent reunification in 1990. We thus exclude East-German counties from the first period of analysis,[16] which gives us 730 ($= (408-86)+408$) observations

[15] We choose the same countries as Dauth et al. (2014) to instrument German imports and exports: Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the United Kingdom. This set of countries excludes Eurozone countries because their demand- and supply conditions are likely correlated with Germany's. See Dauth et al. (2014) for a discussion of this selection.

[16] If we let the first period begin with first democratic election in East Germany in 1990, we could not observe many of the small parties we observe otherwise, since it took time for them to build up party organizations in the East. Thus, a 1990-1998 comparison is for East German districts more or the less equivalent to a 1998 cross-sectional analysis for most but the major parties. Moreover, it took time to privatize the state-owned enterprises dominating the East German economy, casting doubt on the reliability of our measure of trade exposure for East German labor markets in period 1. Since Berlin cannot unambiguously be classified as East or West, we drop this city state from the sample. To

Table 2: The Core Variables in 1987–1998 and in 1998–2009

	(1)	(2)	(3)	(4)	(5)	(6)
	Period 1 (1987-1998), N=322			Period 2 (1998-2009), N=408		
percentile:	25th	median	75th	25th	median	75th
Regressors:						
T_{it}	-0.264	0.068	0.521	-1.222	-0.663	-0.144
instrumented T_{it}	-0.068	0.143	0.402	-1.150	-0.574	-0.113
M_{it} *(Labor Market Outcomes):*						
Δ Share Manufacturing Employment	-4.505	-2.686	-0.987	-1.732	-0.711	0.593
Δ log(Mean Manufacturing Wage)	0.104	0.122	0.147	-0.008	0.022	0.051
Δ log(Mean Non-Manufacturing Wage)	0.086	0.102	0.117	-0.093	-0.071	-0.046
Δ log(Total Employment)	-0.067	0.001	0.081	-0.110	-0.044	0.021
Δ Share Unemployment	0.492	1.259	1.983	-2.138	-1.234	-0.650
Δ log(Total Pop)	0.058	0.099	0.133	-0.046	0.000	0.033
Y_{it} *(Voting Outcomes):*						
Δ Turnout	-0.034	-0.020	-0.012	-0.167	-0.128	-0.095
Δ Vote Share CDU/CSU	-9.234	-7.659	-5.730	-4.493	-2.258	0.620
Δ Vote Share SPD	4.120	6.472	8.248	-19.904	-17.936	-16.079
Δ Vote Share FDP	-2.933	-2.188	-1.467	6.942	8.459	9.820
Δ Vote Share Green Party	-1.779	-1.282	-0.616	2.513	3.673	4.770
Δ Vote Share Extreme-Right Parties	1.520	2.086	3.099	-1.525	-1.021	-0.478
Δ Vote Share Far-Left Parties	0.677	0.908	1.165	5.688	7.078	8.373
Δ Vote Share Small Parties	1.211	1.487	1.796	0.716	1.514	2.525

Notes: Period one (1987–1998) is for West German labor markets only, N = 322. Period two (1998–2009) is for West plus East German labor markets, N = 408. The numbers for 1998–2009 do not change substantively if we drop the East. The table displays the 25th percentile, median, and 75th percentile of three sets of variables: regressors, voting outcomes, and economic outcomes.

in total. Table 2 provides descriptive statistics for our main variables. The table is organized in the following way: Each row presents the distribution of one variable, sliced into its 25th percentile, median, and 75th percentile. Columns 1–3 do this for Period 1 from 1987–1998, and columns 4–6 for Period 2 from 1998–2009. T_{it} is defined in units of 1,000 € per worker in constant 2005 prices.

A comparison of columns 1–3 and 4–6 shows that trade exposure was relatively balanced between import competition and export access in Period 1, with an average T_{it} of just 68 € per worker. In Period 2, trade exposure was more export-heavy, with changes in export access exceeding changes in import competition by on average 663 € per worker.[17] Online Appendix A Figure 1 illustrates the spatial dispersion of the net exposure measure in our data.

Looking at the labor market outcomes, we find evidence of economic stagnation in Period 1. Most importantly, we see a decline in the share of manufacturing employment across all regions concurrent with increasing unemployment. Indeed, Germany was considered "the sick man of Europe" during the 1990s. The period of stagnation was followed by an equally prolonged export and productivity boom. Following Gerhard Schröder's electoral victory in 1998, Germany's inflexible labor market institutions underwent substantial reforms, see (Dustmann et al., 2014). In the course of these reforms, we observe important changes in the behavior of trade unions and employers' associations. Wage policies became more moderate and firms and local labor union chapters were now allowed to deviate from collective bargaining agreements to flexibly adopt to local labor market conditions.[18] As a result of these reforms, the decline in manufacturing employment slowed down and unemployment decreased during Period 2. Furthermore, we observe more moderate or even negative wage growth in this period.

Finally, the table shows substantial variation in political trends across the two periods. From 1987 to 1998, established parties saw an average 4.7 percentage point reduction in their share of the popular vote, while small parties and the extreme right saw an increasing vote share. From

homogenize the sample, we also drop the other two city states, Hamburg and Bremen.

[17] Dauth et al. (2014) explore this finding in detail, and show that trade exposure with Eastern Europe –the dominant shock in period one– was primarily associated with intra-industry trade in final products, i.e., Eastern European final products displaced German final products in German markets. By contrast, trade with China–which was more dominant in period two–was primarily inter-industry, i.e., Chinese imports displaced imports from other countries rather than German production.

[18] A perusal of the *OECD Labour Market Policies and Institutions Indicators Database* nicely illustrates this regulatory change. On the core 'strictness of employment protection' index, Germany stayed in a tight band between 3.13–3.25 throughout Period 1, but this measure then dropped rapidly to an average of 1.46 during Period 2. See www.oecd.org/employment/emp/employmentdatabase-labourmarketpoliciesandinstitutions.htm

1998 to 2009, the main parties CDU and SPD as well as the extreme-right parties lost electoral support.[19]

In summary, 1987–1998 saw changes in import competition and export access that roughly balanced out, economic stagnation and an increase in support for the extreme right. This was followed by increased export access, economic stabilization, and political moderation in period 2. Period-by-region fixed effects will largely absorb these secular trends in our empirical analyses, as well as accounting for the unbalanced panel that arises from not considering East German counties in Period 1.

3 Baseline Results

Our empirical analysis is organized in the following way. Section 3.1 evaluates the total causal effect of trade exposure (T_{it}) on voting behavior (Y_{it}). Section 3.1.1 explores this effect in more detail using individual level data from the SOEP. Section 3.2 examines the effect of trade exposure on six labor market variables, and aggregates this information into principal components (M_{it}).

3.1 Estimating the Total Effect of T (Trade Exposure) on Y (Voting)

We estimate the following Second Stage equation using TSLS:

$$Y_{it} = \Gamma_T^Y \cdot T_{it} + \Gamma_X^Y \cdot X_{it} + \epsilon_{it}^Y \tag{3}$$

Treatment T_{it} represents trade exposure as defined in (1). Outcome Y_{it} denotes changes in voting behavior in county i over period t. Specifically, these are changes in turnout, and changes in the vote-shares of incumbent, small, extreme-right, and far-left parties. Together, these vote shares cover the entire political spectrum. Γ_T^Y is our estimator for the total effect of trade exposure on voting, see (25). X_{it} denotes a selection of control variables. These are i's start-of-period manufacturing employment share; the start-of-period employment share that is college educated, foreign born, or female; the employment share in the largest sector;[20] along with separate controls

[19] The large decrease in SPD vote share reflects the party breaking with its left wing, which subsequently merged with the socialist party PDS to form the new party *Die Linke*. In our data, *Die Linke* is classified as far left. See section Online Appendix B for more details.

[20] It is a feature of the German economy that some regions are dominated by one specific industry. In such regions,

for the employment share in car manufacturing and the chemical industry;[21] start-of-period vote-shares for all parties; voter turnout, start-of-period unemployment rate, and population-share of retirement age.

$\Gamma_X^Y \cdot X_{it}$ further includes a set of period-specific region fixed effects (North, West, South, and East Germany) with the regions being comparable to U.S. Census divisions (Dauth et al., 2014).[22] The regional fixed effects are time-varying to allow for different trends in voting behavior over the periods 1987–1998 and 1998–2009, as evident from the descriptive statistics in section 2.5.

The First Stage equation is

$$T_{it} = \Gamma_{IM}^{T} \cdot Z_{it}^{IM} + \Gamma_{\text{EX}}^{T} \cdot Z_{it}^{EX} + \Gamma_{X}^{T} \cdot X_{it} + \epsilon_{it}^{T}, . \tag{4}$$

Instruments Z_{it}^{IM} and Z_{it}^{EX} are defined in (2). Control variables X_{it} are the same in the first and second stage. Standard errors ϵ_{it} are clustered at the level of 96 commuting zones defined by the Federal Office for Building and Regional Planning (BBR). Table 3 presents our baseline results. Each cell reports results from a different regression. Rows specify different outcome variables, and columns refer to different regression specifications. Results for the coefficients on all control variables are reported in Online Appendix C (table 1).

In our least conservative specification (column 1 of table 3), we consider the start-of-period manufacturing employment share as the only control. We always control for a region's start-of-period manufacturing share in employment because it inherently drives part of the variation in T_{it}; see the discussion in 2.4. In column 2, we add controls for the structure of the workforce, i.e., the start-of-period employment share that is college educated, foreign born, or female. In column 3, we account for the disproportionate regional employment share of some firms by including a control for the employment share in the largest sector, along with separate controls for the employment share in car manufacturing and the chemical industry. In column 4, we add start-of-period vote-shares for all party outcomes and turnout. Finally, in column 5, we add the start-of-period unemployment rate and the population-share of retirement age. This is the most conservative specification, and our preferred one. In this specification, a one-standard-deviation increase in T_{it}

individual firms (e.g. Daimler-Benz, Volkswagen, or Bayer) are likely to have political bargaining power, and as a result politicians may help buffer trade shocks to limit adverse employment effects.

[21] The latter account for those industries' outstanding importance for the German economy.

[22] Each of Germany's 16 states (*Bundesländer*) is fully contained inside one of these four regions.

(1,350 €) increases the extreme-right vote share by 0.12 ($0.09 \cdot 1.35$) percentage points, roughly 28 percent of the average per-decade increase of 0.43 percentage points during the 22 years we study. Column 6 reports the results from our preferred specification as beta coefficients to facilitate comparison between the effects on election outcomes.

The effects are broadly consistent across all five specifications, though we see that the stepwise inclusion of controls reduces the effect size. Our findings suggest no effect on turnout; and looking at reactions across the political spectrum, we see no significant effects on established, small, or far-left parties in our preferred specification in column 5. The only segment of the party spectrum that responds consistently to trade shocks across all specifications is the vote-share of extreme-right parties.[23] Looking at the beta coefficients reported in column 6, we see that the estimated effects for all parties except the extreme right are not only insignificant but also small compared to the effect on extreme-right parties. For a better understanding of potential biases, we present corresponding OLS estimates in table 2 of Online Appendix C. A comparison between IV and OLS estimates for the effect on extreme-right parties shows that the OLS coefficient is consistently smaller than the IV coefficient. This result is in line with our concern that trade exposure partly reflects domestic sectoral demand shifts.[24] We also estimate results based on gravity residuals. This approach does not use IV but instead estimates the exogenous evolution of industry-specific Chinese and Eastern European comparative advantage over Germany based on a comparison of bilateral trade flows of Germany and 'China plus Eastern Europe' vis-a-vis the same set of destination markets that the IV approach uses.[25]. The gravity results are reported in Appendix A and Online Appendix D and are in line with those in table 3.

Overall, the estimated total effect of trade exposure on support for the extreme right is small in absolute terms. To some extent, this small coefficient may be due to measurement error. This is

[23] However, the coefficient for the market-liberal FDP shows a marginally insignificant t-statistic of 1.58, and for turnout we see a t-statistic of 1.22. The latter indicates that turnout might increase with trade exposure. This would complement Charles and Stephens (2013), who find that positive economic shocks decrease voter turnout. One possible explanation for the positive though marginally insignificant effect on votes for the liberal FDP is that regions hit by a trade shock may face increasing demand for redistribution or government intervention in markets (Rodrik, 1995). As a result, those who do not approve such policies may choose to vote for the FDP. Based on our reading of German politics, we take this as a hint for possible polarization, if the economically liberal FDP became an attractive choice for voters who position themselves against growing anti-globalization sentiments in their region.

[24] For example, booming domestic production may increase demand for intermediate input imports, but this is unlikely to have the same political consequences as import competition.

[25] See Autor et al. (2013) and Dauth et al. (2014) for a discussion of the gravity residuals approach relative to the IV approach

Table 3: Effect of Trade Exposure (T_{it}) on Voting (Y_{it})

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	+ Structure	+ Industry	+ Voting	+Socio	Standard.
	IV	IV	IV	IV	IV	IV
Δ Turnout	0.002 (0.939)	0.003 (1.192)	0.004 (1.455)	0.002 (1.095)	0.002 (1.223)	0.036 (1.223)
Established Parties:						
Δ Vote Share CDU/CSU	-0.128 (-0.744)	-0.130 (-0.808)	-0.180 (-0.993)	-0.062 (-0.475)	-0.066 (-0.501)	-0.016 (-0.501)
Δ Vote Share SPD	-0.020 (-0.129)	0.004 (0.030)	-0.006 (-0.039)	-0.011 (-0.090)	-0.009 (-0.073)	-0.001 (-0.073)
Δ Vote Share FDP	0.215*** (2.788)	0.176** (2.384)	0.170** (2.197)	0.109 (1.377)	0.119 (1.583)	0.022 (1.583)
Δ Vote Share Green Party	-0.132** (-2.294)	-0.055 (-1.309)	-0.030 (-0.612)	-0.025 (-0.551)	-0.018 (-0.413)	-0.006 (-0.413)
Non-established Parties						
Δ Vote Share Extreme-Right Parties	0.118*** (3.370)	0.099*** (3.118)	0.113*** (2.845)	0.086** (1.980)	0.089** (2.055)	0.044** (2.055)
Δ Vote Share Far-Left Parties	-0.037 (-0.289)	-0.078 (-0.643)	-0.080 (-0.639)	-0.068 (-0.588)	-0.092 (-0.859)	-0.024 (-0.859)
Δ Vote Share Other Small Parties	-0.015 (-0.391)	-0.017 (-0.458)	0.013 (0.327)	-0.028 (-0.687)	-0.024 (-0.564)	-0.018 (-0.564)
First Stage:						
Z^{IM}_{it}	0.225*** (8.220)	0.234*** (8.350)	0.221*** (7.816)	0.220*** (7.966)	0.220*** (7.971)	0.220*** (7.971)
Z^{EX}_{it}	-0.211*** (-8.519)	-0.212*** (-8.251)	-0.208*** (-8.065)	-0.201*** (-7.660)	-0.202*** (-7.568)	-0.202*** (-7.568)
F-Stat. of excluded Instruments	43.81	43.64	40.15	38.77	38.21	38.21
Period-by-region F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: (*a*) Each cell reports results from a separate instrumental variable regression. The data is a stacked panel of first-differences at the *Landkreis* level. Each regression has 730 observations, i.e. 322 *Landkreise* in West Germany, observed in 1987–1998 and 1998–2009, and 86 *Landkreise* in East Germany, observed only in 1998–2009. We drop three city-states (Hamburg, Bremen, and Berlin in the East). (*b*) All specifications include region-by-period fixed effects. Column 1 controls only for start-of-period manufacturing. Column 2 adds controls for the structure of the workforce (share female, foreign, and high-skilled). Column 3 adds controls for dominant industries (employment share of the largest industry, in automobiles, and chemicals). Column 4 adds start-of-period voting controls. Column 5 is our preferred specification, adding start-of-period socioeconomic controls (population share unemployed, and individuals aged 65+). Finally, Column 6 presents our preferred specification with standardized outcome variables to facilitate comparison. (*c*) The bottom panel presents the first stage results. It reports coefficients for only the two instruments, but includes the full set of controls from the top-panel. (*d*) Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1.

because election results are observed at the *place (Landkreis) of residence* while the labor market data that we use to calculate T_{it} are reported at the *place (Landkreis) of work*. The imperfect overlap of the spatial units is likely attenuating our estimates toward zero. From 1999 on, labor market data are reported at the place of work *and* the place of residence. In Online Appendix C.2, we replicate our main result for the period 1999-2009 with *place-of-residence* data only to gauge the attenuation from combining *place-of-residence* voting data with *place-of-work* employment data. We find that the trade effect on the extreme-right is around 50 percent larger when we use *place-of-residence* labor market data, see Online Appendix C.2 (table 3). However, even a 50 percent larger effect remains small in absolute terms. The plausible explanation for this finding is that Germany did not have a populist leader (or party) with broad appeal like Marine Le Pen in France, Nigel Farage in the UK, or Donald Trump in the U.S. during our study period. All anti-globalization parties at the right fringe were extremist parties with neo-Nazi ties and associations to the *Third Reich* which made them anathema to most Germans. The coefficient size is thus specific to the political context, and our focus is therefore not on the magnitude of the effect of trade exposure on voting behavior but on the causal mechanisms underlying it.

3.1.1 Individual-Level Evidence

In this section, we test whether our regional-level results can be confirmed at the individual level. We do this using the SOEP, an annual household survey that started in 1984 (GSOEP, 2007). The SOEP is unique amongst attitudinal socio-economic surveys in its long-run panel structure.[26] Importantly, we observe individuals in local labor markets. As a result, we can associate individual workers w with their local labor market i's trade exposure (T), instrument T with Z as before, and add the same set of regional controls.[27] This allows us to track decadal changes in individuals' party preferences in a way that mirrors our main local labor market analysis.[28] In addition, we can control for individual characteristics including age, educational attainment, and gender. For our purpose, the relevant GSOEP question asks: *"If there was an election today, who would you vote for?"* We translate this question into a series of dummies that reflect the full party spectrum also

[26] The General Social Survey for example only added a panel component in 2008.

[27] We also face the same attenuation bias as before, with trade exposure being measured at the place of work but individual voting intentions at the place of residence.

[28] Because the SOEP only started to ask about voting intentions for the full party spectrum in 1990 we use the time windows 1990-1998 and 1998-2009, which implies a slightly shorter Period 1 compared to our main results.

observed in table 3, e.g. one dummy if the individual would you vote for the CDU, one if the individual would vote for the SPD, etc.[29] For each party, we aggregate individuals' self-reported voting intentions into a decadal cumulative share of years in which a respondent answered in the affirmative. Based on this, we calculate Y^P_{wt}, as the ratio of the number of years that w states a preference for party P, divided by the number of years that w answered the question in the SOEP.

It is better to measure the outcome as a cumulative share for the whole period instead of using a first difference approach because the latter relies only on individuals' answer at the beginning and the end of the period. Moreover, respondents do not answer all questions in every year, which would increase the number of missing observations in a first difference specification. By contrast, with a cumulative share we simply sum up the instances in which a question was answered in the affirmative and divide by the number of years where we observe an answer. As a result, we obtain about three times as many 'person-decade' observations using the share measure than with the first-difference measure. For each party P, the dependent variable is a share between 0 and 1 for individual w in time period t and we separately estimate

$$Y^P_{wt} = \gamma^Y_{Y-1} \cdot Y^P_{wt-1} + \gamma^Y_T \cdot T_{it} + \gamma^Y_X \cdot X_{it-1} + \epsilon_{wt}. \tag{5}$$

for each party outcome.

With a slight abuse of notation, Y^P_{wt-1} controls for w's survey response to the same question in the base year. X_{it} refers to the same set of regional controls for the base-year as in table 3. Our focus is on estimating γ^Y_T, the effect of region i's trade exposure T_{it} on a resident worker w's reported party support.

Table 4 reports the results. Across rows it mimics closely our main table 3, except that there is no turnout measure in the SOEP. Every coefficient in table 4 reports the estimate of γ^Y_T from a separate regression. T_{it} is always instrumented as before, although we do not report the first stage regressions again. Column 1 includes period and regional fixed effects as well as the regional economic controls from table 3, the most important one of which is a region's baseline manufacturing employment share. We also add region i's base-year socio-economic and voting controls X_{it-1} from table 3 for each period. To better gauge magnitudes, column 2 reports the same specification

[29] There is no question on turnout in the SOEP.

Table 4: Individual-Level Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All Controls	Standardized	1990-1998	1998-2009	High-Skill	Low-Skill & Manuf.	Low-Skill & Not Manuf.	Low-Skill & Manuf., 1998-2009	Low-Skill & Not Manuf., 1998-2009
Established Parties:									
Would Vote CDU/CSU	0.001 (0.292)	0.003 (0.292)	-0.025 (-0.743)	0.002 (0.227)	-0.007 (-0.278)	-0.013 (-0.794)	0.008 (0.827)	-0.006 (-0.350)	0.003 (0.257)
Would Vote SPD	-0.008* (-1.901)	-0.016* (-1.901)	0.027 (0.761)	-0.019** (-2.217)	-0.013 (-0.460)	-0.011 (-0.400)	-0.017* (-1.930)	0.001 (0.031)	-0.022** (-2.352)
Would Vote FDP	0.001 (0.459)	0.005 (0.459)	-0.038 (-0.720)	0.015 (1.177)	-0.018 (-0.420)	0.011 (0.664)	0.007 (0.568)	0.002 (0.116)	0.021 (1.431)
Would Vote Green Party	0.003 (1.000)	0.012 (1.000)	0.019 (0.409)	0.016 (1.295)	0.070 (1.474)	0.025 (0.909)	0.002 (0.152)	0.007 (0.363)	0.007 (0.565)
Non-Established Parties:									
Would Vote Extreme-Right Parties	0.003 (1.619)	0.023 (1.619)	0.029 (0.735)	0.028* (1.802)	0.010 (0.875)	0.083** (2.206)	0.006 (0.475)	0.088** (2.013)	0.016 (1.035)
Would Vote Far-Left Parties	-0.001 (-1.059)	-0.007 (-1.059)	-0.008 (-0.670)	-0.005 (-0.751)	-0.051 (-1.358)	0.019 (1.356)	-0.009 (-1.055)	0.026 (1.579)	-0.010 (-1.043)
Would Vote Other Small Parties	-0.001 (-0.642)	-0.007 (-0.642)	0.018 (0.340)	-0.012 (-1.072)	0.005 (0.182)	-0.026 (-1.053)	-0.003 (-0.190)	-0.048* (-1.674)	-0.001 (-0.042)
Period-by-region F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,669	9,669	3,694	5,975	1,348	2,199	6,122	1,168	3,817

Notes: (*a*) Each cell in this table reports on a separate regression. An observation is an individual w over a period t, where we consider 1990–1998, and 1998–2009, closely mirroring the local labor market results. Each row reports on survey responses to a different question. For every question/outcome y the left-hand-side variable is the share $\frac{\text{\# years that } w \text{ would vote for party } y}{\text{\# years } w \text{ answered the question}}_{it}$. The reported coefficient in all cells is the IV coefficient of regional trade exposure T_{it}. (*b*) Column 1 is the baseline specification which includes period and four region fixed effects as well as all the regional economic, voting and demographic controls from table 3, and individuals' base-year stated political preferences. This is the full set of controls included in all columns. To better gauge magnitudes, columns 2–9 standardize all outcomes by their mean. Columns 3–4 split the sample by period (3,694 + 5,975= 9,669). The results are driven entirely by period 2, i.e. after Germany's labor markets were de-regulated. No part of the political spectrum responds in period 1. In period 2, SPD support is reduced in response to trade exposure and support for the extreme right goes up. In columns 5–7, we break the sample by individuals' skill as well as by whether they are employed in the manufacturing sector (1,348 + 2,199 + 6,122 = = 9,669). High-skill workers (column 5) do not appear to change their political support at all in response to trade exposure. Column 6 shows that it is the population most affected by trade exposure – low-skill manufacturing workers – that drives the effects on the far right. Interestingly, the mainstream-left results are driven by low-skill service workers in column 7. We conjecture that this may be because they experience increased competition from laid-off low-skill manufacturing workers. In columns 8–9, we focus on the second period, which again sharpens the results from columns 6–7. (*c*) Standard errors are clustered at the region level. *** p<0.01, ** p<0.05, * p<0.1.

with standardized outcomes.

Regional trade exposure shifts individuals' preferences to the extreme right, though the effect is marginally insignificant with a t-stat of 1.619. In the individual-level data there is stronger evidence of a reduction in preference for the established left-wing party, the SPD, which comes out much less clearly in the aggregate results in table 3. No other party across the entire spectrum shows a response that is close to being significant. The discussion in section 2.5 suggests that the effect of trade shocks on labor markets should be more pronounced in the second period, when companies were more flexible to react. We therefore report the results separately by period in columns 3 and 4. It turns out that both the extreme right and SPD results are driven entirely by period 2, i.e. after Germany's labor markets were de-regulated. We will also find this at the regional level, see table 6.

Once we dig deeper into what types of workers are driving the observed patterns we find distinctive results. In columns 5–7 we split the sample by skill as well as by whether an individual works in manufacturing, i.e. whether their employment sector is more heavily exposed to trade competition.[30] Both the extreme right effect and the SPD effect are entirely driven by low-skill workers, while high-skill workers do not respond at all.[31] Splitting the low-skill sample into manufacturing and non-manufacturing employment, we see that the extreme-right response is entirely driven by low-skill workers in manufacturing sectors, i.e. those likely to be most exposed to competition from low-wage countries. For this subpopulation the effect is also much larger. By contrast and most interestingly, the reduction in the change in the SPD's vote share is entirely driven by low-skill *non*-manufacturing workers. A possible interpretation is that low-skill workers in the service sector are affected by competition from laid-off manufacturing workers, or that laid-off manufacturing workers had to accept unattractive jobs in the service sector. In either case, they might blame the SPD-induced labor market reforms, such that trade exposure would only indirectly affect their changing party support. Columns 8–9 show that this pattern is again driven by the second period. In summary, the individual level evidence confirms our main findings

[30] In an earlier working paper, we focused on comparing the effect of individuals' trade exposure due to their industry of employment relative to their regions' trade exposure (Dippel, Gold, and Heblich, 2015). However, we have come to the conclusion that individuals' industry of employment is measured too coarsely in the SOEP to draw strong conclusions about the relative importance of these two types of trade exposure.

[31] The SOEP reports skills as educational attainment according to the 'ISCED-1997' classification, where 'high' means some college.

of a political backlash to increasing international trade, and additionally shows that those who are most likely to experience adverse labor market effects are the ones who respond the most to increasing trade exposure.

3.2 The Effect of Trade Exposure T on Labor Markets M

We now turn to estimating equation

$$M_{it} = \Gamma_T^M \cdot T_{it} + \Gamma_X^M \cdot X_{it} + \epsilon_{it}^M, \tag{6}$$

again using TSLS.

We observe six separate labor market outcomes, i.e. manufacturing's employment share, total employment, manufacturing and non-manufacturing wages, unemployment, and total population size; most of which have been found to be affected by trade exposure in the existing literature already. However, we are more interested in the extent to which labor market as a whole act as causal mechanism for the causal relationship between trade exposure and voting behavior, than in the partial effects of single variables. To this end, we develop an analytic framework in section 4 that focuses on identifying a single mechanism's mediating effect. Therefore, we must reduce the dimensionality of the labor market data to obtain a concise measure of M_{it}. A natural way of doing so is to conduct a principal component analysis (PCA). PCA is appealing for two reasons: First, PCA combines all observed labor market variables into aggregated 'labor market components' (LMC) that condense labor market disturbances into their key characteristics. Second, these LMCs are by construction orthogonal so that they can be investigated separately one at a time. PCA results for our six labor market outcomes are reported in table 5. Following the standard "Kaiser-Guttman" criterion, we analyze labor market components with an eigenvalue larger than 1. In our data, the second labor market component (LMC$_2$) has an eigenvalue of 1.415 followed by a big drop in the third's eigenvalue to 0.6085. We interpret this as natural break and consider the first two LMCs in our analysis. Together, LMC$_1$ and LMC$_2$ explain about 80 percent of the variation in the labor market data.

As statistical constructs, principal components are best interpreted through the lens of their factor loadings, which indicate how strongly every labor market outcome relates to each LMC.

Table 5: Principal Component Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Principal Components		Factor-Loadings					
	Eigen-value	Eigen-value: Proportion	Δ Share Manuf. Empl.	Δlog(Avg. Manuf. Wage)	Δ log(Avg. Non-Manuf. Wage)	Δ Log(Total Empl.)	Δ Share Unempl	Δ log(Total Pop)
LMC_1	3.288	0.548	-0.284	0.452	0.484	0.248	0.460	0.456
LMC_2	1.415	0.236	0.606	0.026	-0.098	0.702	-0.215	0.291

Notes: Following the "Kaiser-Guttman" criterion, we retain and analyze principal components ('labor market components') with an eigenvalue above 1. (The third had an eigenvalue of 0.609.) The first column shows the eigenvalues of the first two labor market components. The second column shows the share of total data variation they explain. Together, the two LMCs explain almost 80 percent of the variation in the data $(0.548 + 0.236)$. Moving on to factor loadings, the first LMC is associated with changes in total population and wages, as well as with unemployment. The second LMC is strongly associated with changes in the manufacturing share of employment and in total employment.

The first labor market component's factor loadings are strongly positive for changes in wages, total population, and unemployment. The second labor market component's factor loadings are strongly positive for changes in the share of manufacturing employment and changes in total employment, and negative for changes share in unemployment. Broadly speaking, increases in either LMC's values indicate "positive" developments on the local labor markets, since overall employment increases. However, increasing employment appears to be driven by different factors, and accompanied by different side-effects. An increase in LMC_1 is accompanied by decreasing manufacturing employment and increasing unemployment. On the contrary, increases in LMC_2 are largely driven by increasing manufacturing employment and decreasing unemployment, but less related to wage changes. The urban agglomeration literature offers a plausible interpretation for these factor loadings. Duranton and Puga (2005) point out that regional specialization has become "functional" as opposed to "sectoral" over the last decades, implying a tendency for headquarters and business services to cluster in large cities, a trend that appears to be clearly borne out in Germany (Bade, Laaser, and Soltwedel, 2003). The first labor market component seems to capture this structural change. LMC_1 describes labor markets where job growth is driven by the services industry, with higher-paying jobs being created but lower-skilled manufacturing workers becoming unemployed, and population density increasing. Arguably, this resembles local labor markets in urban regions. Conversely, the second labor market component seems to be much more related to ”classical” manufacturing industries as they dominate local labor markets outside the large cities.[32] The question is whether the aggregate labor market components are causally affected by trade exposure. To facilitate the interpretation an ease comparison with the existing literature, we additionally assess the underlying labor market variables’ responses to trade exposure.

Γ_T^M is our estimator for the effect of trade exposure on the labor market components and their underlying variables, see (24). Trade exposure T_{it}, control variables X_{it}, and instruments Z_{it} in equation (6) are the same as the ones used to estimate the effect of trade on voting behavior in (3). We only replace the political outcomes Y_{it} in (3) with the labor market outcomes M_{it} of (6).

The results are displayed in table 6, which is structured in exactly the same way as table 3. Each cell reports the result from a different regression specification. Column 1 is our baseline specification; column 2 adds structural characteristics of the workforce, i.e., the employment shares of female, foreign, and high-skilled workers; column 3 adds controls for the employment share in the largest industry, along with controls for the employment shares in the automobile and chemical sector; column 4 adds voting controls; and finally, our preferred specification in column 5 also includes socio-economic controls for the unemployment share and the share of individuals over age 65.[33] Column 6 reports the results from our preferred specification as beta coefficients to facilitate comparison with the effects on voting outcomes.

At the top of the Second Stage panel of table 6, we report the effect of T_{it} on the two LMCs measuring labor market disturbances on an aggregate level. It is evident that trade exposure has no bearing on LMC$_1$. This does not come as big surprise, given the factor loadings reported in table 5. Most obviously, the LMC$_1$ strongly depends on the services industry, which is less affected by trade in manufacturing. In contrast, trade exposure turns out to be an important driver of labor market disturbances described by LMC$_2$, which more strongly depends on the manufacturing industries. This conjecture becomes clearer when looking at the effects of trade exposure on single labor market outcomes, and their relation to the labor market components. By now, we can conclude that only LMC$_2$ is a candidate for mediating the effect of trade exposure on

[32] In unreported regressions, we do indeed find LMC$_1$ to be significantly positively correlated with urban centers and negatively correlated with urbanized regions and rural regions. On the contrary, the LMC$_2$ is significantly positively correlated with urbanized regions and negatively correlated with urban centers and rural regions.

[33] In tables 3 and 6, we run two separate two-staged least squares systems that share the same instrument. Because of this, we use the exact same set of controls in both tables, adding some potentially irrelevant social and voting controls to the labor market specifications in columns 4–5 of table 6. As a result of this minor simplification, there are no efficiency gains from estimating the two equations jointly in *seemingly unrelated regressions* (SUR) (Wooldridge, 2002, p. 143-146).

Table 6: Effect of Trade Exposure T_{it} on Labor Markets

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	+ Structure	+ Industry	+ Voting	+Socio	Standard.
	IV	IV	IV	IV	IV	IV
1st Labor Market Component: LMC_1	-0.105**	-0.050	-0.045	-0.032	-0.021	-0.011
	(-2.108)	(-1.313)	(-1.150)	(-0.903)	(-0.679)	(-0.679)
2nd Labor Market Component: LMC_2	-0.265***	-0.301***	-0.328***	-0.324***	-0.322***	-0.271***
	(-2.894)	(-3.526)	(-3.667)	(-3.696)	(-3.755)	(-3.755)
Δ Share Manufacturing Employment	-0.440**	-0.618***	-0.738***	-0.745***	-0.755***	-0.247***
	(-1.979)	(-3.098)	(-3.601)	(-3.677)	(-3.745)	(-3.745)
Δ log(Mean Manufacturing Wage)	-0.006**	-0.005**	-0.006**	-0.005**	-0.006***	-0.083***
	(-2.496)	(-2.145)	(-2.466)	(-2.501)	(-2.592)	(-2.592)
Δ log(Mean Non-Manufacturing Wage)	-0.005***	-0.002*	-0.002	-0.001	-0.001	-0.015
	(-2.864)	(-1.666)	(-1.027)	(-0.785)	(-0.808)	(-0.808)
Δ log(Total Employment)	-0.023***	-0.024***	-0.025***	-0.025***	-0.024***	-0.207***
	(-2.853)	(-3.131)	(-3.203)	(-3.239)	(-3.295)	(-3.295)
Δ Share Unemployment	0.076	0.097	0.076	0.084	0.110*	0.060*
	(1.100)	(1.540)	(0.918)	(1.031)	(1.694)	(1.694)
Δ log(Total Population)	-0.009***	-0.007***	-0.006**	-0.005**	-0.004*	-0.050*
	(-3.108)	(-2.903)	(-2.381)	(-2.254)	(-1.852)	(-1.852)
First Stage:						
Z^{IM}_{it}	0.225***	0.234***	0.221***	0.220***	0.220***	0.220***
	(8.220)	(8.350)	(7.816)	(7.966)	(7.971)	(7.971)
Z^{EX}_{it}	-0.211***	-0.212***	-0.208***	-0.201***	-0.202***	-0.202***
	(-8.519)	(-8.251)	(-8.065)	(-7.660)	(-7.568)	(-7.568)
F-Stat of excluded Instruments	43.81	43.64	40.15	38.77	38.21	38.21
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: (*a*) Each cell reports results from a separate instrumental variable regression. The data is a stacked panel of first-differences at the *Landkreis* level. Each regression has 730 observations, i.e. 322 *Landkreise* in West Germany, observed in 1987–1998 and 1998–2009, and 86 *Landkreise* in East Germany, observed only in 1998–2009. We drop three city-states (Hamburg, Bremen, and Berlin in the East). (*b*) All specifications include region-by-period fixed effects. Column 1 controls only for start-of-period manufacturing. Column 2 adds controls for the structure of the workforce (share female, foreign, and high-skilled). Column 3 adds controls for dominant industries (employment share of the largest industry, in automobiles, and chemicals). Column 4 adds start-of-period voting controls. Column 5 is our preferred specification, adding start-of-period socioeconomic controls (population share unemployed, and individuals aged 65+). Finally, Column 6 presents our preferred specification with standardized outcome variables to facilitate comparison. For outcomes in logs, the table reports on a semi-elasticity: For example, a one-standard-deviation increase in T_{it} (€1,350) decreased total employment by about 3 percent, ($e^{-0.024 \cdot 1.35} - 1 = -0.032$). (*c*) The bottom panel presents the first stage results. It reports coefficients for only the two instruments, but includes the full set of controls from the top-panel. (*d*) Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1.

voting behavior.

With respect to the single labor market variables, we closely replicate results that have already been established in Autor et al. (2013), Dauth et al. (2014) and Pierce and Schott (2016): Trade exposure has a significantly negative effect on manufacturing employment. In our preferred specification, in column 5, a one-standard-deviation (1,350 €) increase in trade exposure decreases the share of manufacturing employment by around one percentage point, i.e. roughly three-quarters of Germany's average by-decade decrease of 1.3 percent over the period. This strongly relates to a decrease in LMC_2, but leads to a moderate increase in LMC_1. Trade exposure also implies small but significant wage cuts in manufacturing industries. This is in contrast to U.S. data, where import competition appears to depress non-manufacturing wages but not manufacturing wages (Autor et al. 2013, table 7), suggesting more downward wage rigidity in U.S. manufacturing. The negative effect on manufacturing wages is more relevant for LMC_1, where it offsets the effect related to manufacturing employment. For LMC_2 the effects of trade exposure on manufacturing wages is less important but aligned with the effect on manufacturing employment. Moreover, trade exposure increases unemployment, which increases LMC_1 but decreases LMC_2. Furthermore, trade exposure depresses total employment, decreasing both labor market components, but with a much stronger loading on LMC_2. Lastly, there is a small but significant negative effect on total population, which is more relevant for LMC_1.[34] As before, Online Appendix C reports corresponding OLS results (table 4), and the coefficients of all controls (table 5). The OLS results are across the board smaller than the IV results, suggesting that industry-specific productivity improvements affecting the OLS estimates were on average labor-saving. Subsequently, we will investigate labor market disturbances as causal mechanism for explaining the causal effect of trade exposure on voting behavior, with a strong focus on the mediating effect of LMC_2.

The discussion in section 2.5 suggests that the effect of trade shocks on labor markets should be more pronounced in the second period when there was less regulation preventing market responses. Having found some evidence for this in the individual results in table 4, we also decompose the main results in tables 3 and 6 by period. To conserve space, we report the results in

[34] One might be concerned that the trade effects on extreme right party support could be driven by selective outmigration of trade exposed individuals. However, the population effect is small ($e^{-0.004 \cdot 1.35} - 1 = -0.005$), and the previous section's individual analysis confirms that individuals exposed to trade to indeed change their voting behavior.

Online Appendix E. Both voting and labor market results were considerably stronger in Period 2 (1998–2009). This symmetry in the response of different outcomes suggests the important role of labor markets as mediators in the transmission from trade shocks to voting responses, but to actually estimate how important we need to now introduce our mediation model.

4 Merging IV Models into a Single Mediation Model

In Section 3, we evaluate two causal parameters: the total effect of trade exposure T on labor markets M and the total effect of T on voting behavior Y. We identify these effects using a standard IV model where the instrument Z is the trade exposure of countries other than Germany as in Autor et al. (2013). This evaluation does not identify the causal effect of M on Y. Indeed the standard IV model is not suitable to examine the causal relation between M and Y. In this section, we modify the simple IV model into a mediation model that enable us to identify the causal effect of M on Y.

Some notation in needed to clarify ideas. We use $supp(T), supp(M)$ for the support of variables T and M. We use $M(t)$ and $Y(t)$ for the potential outcomes of M, Y when T is fixed at value $t \in supp(T)$ and $Y(m)$ for the potential outcome of Y when M is fixed at $m \in supp(M)$. $Y(t, m)$ stands for the counterfactual outcome Y when T is fixed at value $t \in supp(T)$ and M is fixed at value $m \in supp(M)$. We use $\perp\!\!\!\perp$ for statistical independence and $\not\perp\!\!\!\perp$ for its negation. For sake of notational simplicity, we suppress conditioning variables X that we wish to control for. Our analysis can be understood as conditioned on those without loss of generality.

The identification of causal effects of T on M, Y of Section 3 arises from three properties of the instrumental variable:[35]

$$\text{Exclusion Restriction of Labor Market Variables: } Z \perp\!\!\!\perp M(t) \tag{7}$$

$$\text{Exclusion Restriction of Voting Outcomes: } Z \perp\!\!\!\perp Y(t) \tag{8}$$

$$\text{IV Relevance : } Z \not\perp\!\!\!\perp T, \tag{9}$$

The general causal model that generates the IV properties (7)–(9) is given by:[36]

$$T = f_T(Z, V, \epsilon_T), \quad M = f_M(T, V, \epsilon_M), \quad Y = f_Y(T, V, \epsilon_Y), \tag{10}$$

$$\text{where: } Z, V, \epsilon_T, \epsilon_M, \epsilon_Y \text{ are mutually statistically independent.} \tag{11}$$

Model (10)–(11) consists of the four observed variables Z, T, M, Y, three exogenous error terms $\epsilon_T, \epsilon_M, \epsilon_Y$, and an *unobserved confounding variable* V that is the source of endogeneity. Equations (10) define the causal relation among variables: Z causes T, T causes M and Y, and the confounder V causes T, M, Y but not Z. there are no restrictions on the functional forms of f_T, f_M, f_Y in (10).

As mentioned, Model (10)–(11) makes no causal distinction between M and Y. Thus we can interpret Model (10)–(11) as two separated IV models described in Table 7. Our task is to merge these two IV models into a single *mediation model* that seizes the information that M causes Y.[37]

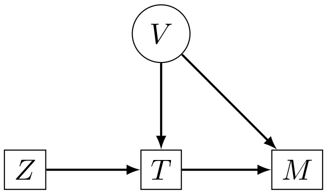
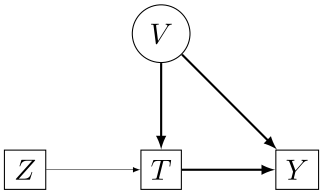
A meditation model enables to unravel the mechanism though which T causes Y. If the causal effect of M on Y were identified, then the total effect of T on Y could be expressed as the sum of the effect of T on Y that operates though the causal chain $T \rightarrow M \rightarrow Y$ (the indirect effect) and

[35] The IV properties of exclusion restriction (e.g. (7)–(8)) and IV-relevance (e.g. (9)) are necessary but not sufficient to identify the causal effect of T on an outcome Y. An extensive literature exists on the additional assumptions that render the identification of treatment effects. For example, if T and Z are continuous and we assume linearity, then causal effects can be evaluated by two-stage least squares. Imbens and Angrist (1994) study a binary T and assume a monotonicity criteria that identifies the Local Average Treatment Effect ($LATE$). Vytlacil (2006) studies categorical treatments T and evoke a separability condition of the choice function. Heckman and Pinto (2017) present a monotonicity condition that applies to unordered choice models with multiple treatments. Pinto (2015) investigate identifying assumptions generated by revealed preference analysis. Heckman and Vytlacil (2005) investigate the binary treatment, continuous instruments and assume that the treatment assignment is characterized by a threshold-crossing function. **?** assume a generalized set of threshold-crossing rules. Imbens and Newey (2007); Blundell and Powell (2003, 2004); Altonji and Matzkin (2005) study control function methods characterised by functional form assumptions.

[36] See Heckman and Pinto (2015a) for a discussion on the causal relations of the standard IV model.

[37] See Online Appendix F for a concise introduction to mediation models. See Appendix Online Appendix G for a discussion on assumptions that are commonly evoked in the literature of mediation analysis.

Table 7: Two Standard IV Models for Labor M and Voting Y

DAG for Labor M	*DAG for Voting Y*
	
Respective IV Model for M	*Respective IV Model for Y*
$T = f_T(Z, V, \epsilon_T)$ $M = f_M(T, V, \epsilon_M)$ $Z \perp\!\!\!\perp V \perp\!\!\!\perp \epsilon_T \perp\!\!\!\perp \epsilon_M$	$T = f_T(Z, V, \epsilon_T)$ $Y = f_Y(T, V, \epsilon_Y)$ $Z \perp\!\!\!\perp V \perp\!\!\!\perp \epsilon_T \perp\!\!\!\perp \epsilon_Y$

the causal effect of T on Y that is not mediated by M (the direct effect):

$$\underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}} = \underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}}. \tag{12}$$

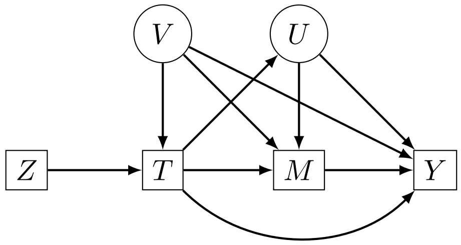
A general mediation model should allow for two sources endogenous effects: an *unobserved confounder* V that causes T, M, Y and an *unobserved mediator* U that is caused by T and causes M, Y. Panel A of Table **??** represents the general mediation model with instrumental variables as a DAG. Panel B describes the model equations.[38]

The counterfactuals $M(t), Y(t)$ of the general mediation model of Table **??** are given by Equations (13)–(14). The unobserved confounder V induces a correlation between T and these counterfactuals, thus the independence relation $T \perp\!\!\!\perp \big(Y(t), M(t)\big)$ does not hold. Otherwise stated, T is endogenous due to V. Nevertheless, $V \perp\!\!\!\perp Z$, which implies that the exclusion restrictions $Z \perp\!\!\!\perp Y(t)$, and $Z \perp\!\!\!\perp M(t)$ still hold. Counterfactual $Y(m)$ in (15) refers to the causal effect of M on Y. Confounder V induces a correlation between M and $Y(m)$ thereby M is endogenous. Moreover, the IV relevance, $Z \not\perp\!\!\!\perp T$, implies that the exclusion restriction $Y(m) \perp\!\!\!\perp Z$ does not hold. In other words, instrument Z does not render the identification of the causal effect of M on Y.

[38] All the model properties discussed in this section also hold if the unobserved confounder V causes the unobserved mediator U.

Table 8: The General Mediation Model with IV

A. DAG Representation



B. Model Equations

$$
\begin{aligned}
\text{Treatment variable: } T &= f_T(Z, V, \epsilon_T),\\
\text{Unobserved Mediator: } U &= f_U(T, V, \epsilon_U),\\
\text{Observed Mediator: } M &= f_M(T, U, V, \epsilon_M),\\
\text{Outcome: } Y &= f_Y(T, M, U, V, \epsilon_Y),\\
&\text{where: } Z, V, \epsilon_T, \epsilon_M, \epsilon_Y, \epsilon_U \text{ are mutually statistically independent.}
\end{aligned}
$$

$$M(t) = f_M(t, U(t), V, \epsilon_M) = f_M(t, f_U(t, \epsilon_U), V, \epsilon_M), \tag{13}$$

$$Y(t) = f_Y(t, M(t), U(t), V, \epsilon_Y) = f_Y(t, M(t), f_U(t, V, \epsilon_U), V, \epsilon_Y), \tag{14}$$

$$Y(m) = f_Y(T, m, U, V, \epsilon_Y). \tag{15}$$

Additional assumptions are needed to identify mediation effects. Specifically, we seek a general mediation model that enables the identification of counterfactual outcomes $Y(t), M(t), Y(m)$ and $Y(m,t)$. Moreover this mediation model should comply with the seven desired properties listed below:

1. The model allows for confounders and unobserved mediators.
2. T and M are endogenous, that is, $T \not\perp\!\!\!\perp \big(M(t), Y(t)\big)$ and $M \not\perp\!\!\!\perp Y(m,t)$.
3. Instrumental variables Z directly cause T.
4. Model does *not* require a dedicated instrument that directly causes M.
5. Instrument Z is suitable to identify three causal relations: $T \to Y$; $T \to M$; and $M \to Y$.
6. Additional causal assumptions must have a clear empirical interpretation.
7. Additional causal assumption to the general mediation model of Table **??** must be testable.

Properties 1 and 2 simply state that the model must account for potential sources of endogenous effects. Property 3 assures that an instrumental variable that causes T exists. Property 4 state that the model should not rely on additional instruments that target the mediation variable M. The existence of dedicated IV for M (in addition to the IV for T) is unlikely in most empirical settings. Property 5 states that a single set of instruments must enable the identification of the three causal effects of interest. Section 4.1 describe our model assumptions and examine the identification of these causal effects. Property 6 inquires about the interpretation of additional model assumptions that enable the identification of counterfactuals. We provide a detailed interpretation of our model assumptions in Section 4.2. Section 4.3 describes the estimation of causal parameters under linearity. Our assumptions reduce the generality of the mediation model in Table **??**. Property 7 states that these additional assumptions should be testable. We discuss a model specification test in Section 4.4.

4.1 The Restricted Mediation Model with Instrumental Variables

Table 9 describes the *Restricted Mediation Model* that complies with the seven desirable properties listed in Section 4. The restricted mediation model differs from the general mediation model of Table **??** by decomposing the confounder V into two unobserved variables: V_T that causes T, M and V_Y that causes Y, M.

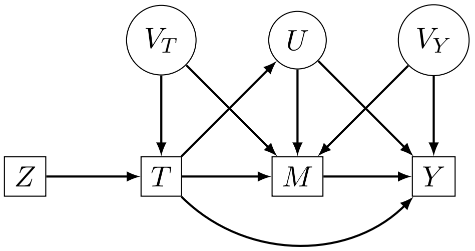
Equations (16)–(19) list the counterfactual outcomes $M(t), Y(t), Y(m), Y(m,t)$ of the restricted model in Table 9:

$$\begin{aligned}
M(t) &= f_M(t, U(t), V_T, V_Y, \epsilon_M), && (16)\\
Y(t) &= f_Y(t, M(t), U(t), V_Y, \epsilon_Y) = f_Y(t, f_M(t, U(t), V_T, V_Y, \epsilon_M), U(t), V_Y, \epsilon_Y), && (17)\\
Y(m) &= f_Y(T, m, U, V_Y, \epsilon_Y), && (18)\\
Y(m,t) &= f_Y(t, m, U(t), V_Y, \epsilon_Y), && (19)\\
\text{where } U(t) &= f_U(t, \epsilon_U).
\end{aligned}$$

Treatment T and Mediator M are endogenous variables in both the general and restricted mediation models. According to (16)–(17), confounder V_T induces a correlation between T and counterfactuals $M(t), Y(t)$, thus T is endogenous and $T \not\perp\!\!\!\perp \big(Y(t), M(t)\big)$. According to (18)–(19), confounder V_Y induces a correlation between M and $Y(m), Y(m,t)$. Thereby M is also endoge-

Table 9: Restricted Mediation Model with IV

A. DAG Representation



B. Model Equations

$$\begin{aligned}
\text{Treatment: } T &= f_T(Z, V_T, \epsilon_T),\\
\text{Unobserved Mediator: } U &= f_U(T, \epsilon_U),\\
\text{Observed Mediator: } M &= f_M(T, U, V_T, V_Y, \epsilon_M),\\
\text{Outcome: } Y &= f_Y(T, M, U, V_Y, \epsilon_Y),\\
\text{Independence: } & V_T, V_Y, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y \text{ are statistically independent.}
\end{aligned}$$

nous and $M \not\perp\!\!\!\perp (Y(m), Y(m,t))$. Nevertheless, the restricted mediation model of Table 9 generates three exclusion restrictions described in Theorem **T-1**.

Theorem T-1 *The following statistical relations hold for the restricted mediation model of Table 9:*

	Targeted Causal Relation	*IV Relevance*		*Exclusion Restrictions*
Property 1	*for* $T \to Y$	$Z \not\perp\!\!\!\perp T$	*and*	$Z \perp\!\!\!\perp Y(t)$
Property 2	*for* $T \to M$	$Z \not\perp\!\!\!\perp T$	*and*	$Z \perp\!\!\!\perp M(t)$
Property 3	*for* $M \to Y$	$Z \not\perp\!\!\!\perp M \vert T$	*and*	$Z \perp\!\!\!\perp Y(m) \vert T$

Proof P-1 *See Appendix B.*

The exclusions restrictions of Properties 1 and 2 in **T-1** are identical to the ones in (7)–(8), generated by the standard IV model of Table 7. These exclusion restrictions arise from the statistical independence between the instrument Z and unobserved confounders V_T, V_Y. Property 3 however states an exclusion restriction that is not generated by the standard IV model, that is, $Z \perp\!\!\!\perp Y(m)|T$. This property implies that instrument Z can be used to evaluate the causal relation of M on Y if (and only if) conditioned on T. Indeed, while $Z \perp\!\!\!\perp Y(m)|T$ holds, $Z \perp\!\!\!\perp Y(m)$ does not.

Properties 1 and 2 of **T-1** allow for the identification of counterfactual outcomes $M(t)$ and $Y(t)$ by applying standard IV techniques. A novel feature of our model is the possibility to use Z to identify the counterfactual outcome $Y(m,t)$. Property 3 of T-1 is useful to identify of the conditional counterfactual $(Y(m)|T=t)$. Corollary C-1 states that $(Y(m)|T=t)$ is equal in distribution to the counterfactual outcome $Y(m,t)$. Therefore Property 3 of T-1 can be used to identify $Y(m,t)$.

Corollary C-1 *In the restricted mediation model of Table 9, the counterfactual outcome $Y(m)$ conditioned on $T=t$ is equal in distribution to the counterfactual outcome $Y(m,t)$, i.e., $(Y(m)|T=t)\overset{d}{=}Y(m,t)$.*

Proof P-2 *See Appendix C.*

4.2 Understanding the Restricted Mediation Model and its Identification

Property 3 of Theorem T-1 may come as a surprise as it states that Z becomes a valid instrument for M conditional on T, despite the fact that Z directly causes T instead of M. Moreover our key identifying assumption is new to the standard training on IV methods. The restricted mediation model imposes a causal restriction on unobserved confounding variables: while confounders directly cause T and M (i.e. V_T), and M and Y (i.e. V_Y), there is no confounder that directly causes T, M and Y jointly. This section clarifies these novel ideas.

Our task is twofold. First we build the intuition on the causal relations of the restricted mediation model of Table 9: *(i)* we interpret the model in light of our empirical context; *(ii)* we explain that our key identifying assumption arises from an economic assessment; and *(iii)* we provide familiar examples in labor economics where the assumption fails. Second, we clarify how the restricted mediation model enables the identification of the causal effect of M on Y : *(i)* we build intuition based a well-known example that uses college distance as an IV to identify the causal effect of college attendance on income; *(ii)* we apply this intuition to examine the identification of causal relations among trade exposure, labor market variations and political polarization.

Model Interpretation

Understanding Confounder V_T: Confounder V_T stands for the unobserved variables that affect trade exposure and labor market variables. The confounder refers to firm-specific decisions.

Firms are hetergeneous in terms of the goods they produce and the industry they belong to. Competition among foreign or domestic products harms firms that firms produce substitute goods while benefit firms that produce complementary goods. The firm-level decision regarding Export/Import and also hiring/firing are affected accordingly.

Understanding Confounder V_Y: Confounder V_Y stands for the unobserved variables that jointly affect labor market outcomes and voting decisions. The confounder is mostly related to worker-specific decisions.

Workers are heterogeneous in terms of their (unobserved) human capital. Fluctuations of the labor market adversely impact the demand of worker's particular set of skills. Some workers may suffer in job insecurity or incur in income losses. These workers are more likely to seek political change towards a protectionist agenda.

Economic Interpretation: The separation of confounders V_T, V_Y is rooted on fact that these confounders refer to different economic agents. In our empirical context, labor market variables M are endogenous in a regression of M on voting Y. Endogenous effects arise due to the heterogeneity of the labor force. Worker unobserved characteristics (e.g. skills) that influence its employability are likely to affect its voting behavior.

Trade exposure T causes voting Y via observed labor market variables M but also though unobserved mediators U. Trade is also endogenous in a regression of voting Y on trade. Endogenous effects are generated by firm heterogeneity. Firms that belong to different industries seeks for workers with distinct skills. Firm unobserved characteristics that affect its trade decisions also affect its hiring decisions. Fluctuations of employment rates of each industry adversely affect the heterogenous labor force, which, in turn, impact voting outcomes.

When Identifying Assumption Fails: The model we present is general and can be potentially applied to a wide range of empirical questions. Nevertheless it is useful to exemplify research questions for which the identifying assumption of the restricted mediation model is unlikely to hold.

Suppose that a researcher is interested in understanding the mechanisms generating earning gaps across college majors. Let earnings be the labor market value of human capital, which comprises three major components: *(i)* unobserved abilities such as cognition; *(ii)* specific knowledge associated with each college major; and *(iii)* academic performance such as the GPA. The distribution of the unobserved cognition (item *(i)*) may differ across majors and it is a source of selection bias. The *total effect* of a college major on earnings is decomposed into the *indirect effect* that is mediated by academic performance (item *(iii)*) and the direct effect that pertain to the major specific knowledge (item *(i)*).

In our notation, T stands for the choice of college major, M for academic performance and Y for earnings. Cognition plays the role of the unobserved confounder V which is likely to affect the major choice T, school performance M and earnings Y. It is unlikely that individual unobserved abilities V could be split into V_T that causes T, M and V_Y that causes M, Y. Therefore the restricted mediation model is not recommended.

Identification and the Role of the Instrumental Variable

Understanding the Identification Result: The novel property of the restricted mediation model is that instrument Z is a valid instrument to identify the causal effect of M on Y when conditioned on T, that is, $Z \perp\!\!\!\perp Y(m)|T$. We use a stylized economic model to clarify the intuition of this result.

Suppose a researcher observes that the income of individuals that perform athletic activities during college is substantially higher than the income of its peers. The researcher is interested in identifying the causal effect of college on income that is mediated by athletic activities.

Consider the stylized model where T is a college indicator, $T = 1$ for college attendance and $T = 0$ otherwise. Outcome Y denotes income during adulthood. Instrument Z stands for the distance between the home of prospective students to college. For sake of simplicity, suppose Z takes two values: $Z = 1$ if college is near and $Z = 0$ if college is far. If $Z \perp\!\!\!\perp (Y(1), Y(0))$ holds and we assume that prospective students are more likely to go to college if it is near, then the Local Average Treatment Effect (LATE) of college attendance on income is identified (see Imbens and Angrist (1994)).

Let the mediator M be the athletic indicator that takes value $M = 1$ if a student enrolls in an athletic club and $M = 0$ otherwise. Suppose that college students walk from their homes to the college. Let the unobserved variable V_T that impacts the college choice T and athletic enrollment M be the size of students legs. Students with long legs ($V_T = 1$) are mode likely to attend college than students with short legs ($V_T = 0$). Long legs ($V_T = 1$) are also desired for athletics ($M = 1$). In short, V_T causes T and M.

College distance Z and leg sizes V_T are statistically independent, that is, $Z \perp\!\!\!\perp V_T$. However, conditioned on going to college ($T = 1$), students that live far from college ($Z = 0$) are more likely to have long legs ($V_T = 1$.) In other words, by conditioning on college attendance ($T = 1$), we induce a negative correlation between leg size V_T and college distance Z. Notationally, we have that $Z \not\perp\!\!\!\perp V_T | T = 1$ even though $Z \perp\!\!\!\perp V_T$.

Relation $Z \not\perp\!\!\!\perp V_T | T$ means that, conditioned on T, variations on Z induce variations on V_T. Effectively it means that college distance Z becomes a proxy for leg size V_T. But leg size V_T causes athletic enrollment M. Thus, conditioned on T, variations on college distance Z induce variations on athletic enrollment M. In other words, Z becomes and instrument for M. At first glance, this result may seem counter-intuitive. Instrument Z directly causes only T and Z is unconditionally independent of V_T. However when T lacks variation (conditioning), then variations in Z induce variations in V_T (induced correlation), which consequently induce variations in M (because V_T causes M).

Trade Interpretation: [RP: This explanation Needs Attention]

Previous literature T is endogenous because there are industry-specific demand trajectories in Germany (V_T). As discussed, the fact that the TSLS estimates in table 6 are larger than the OLS is consistent with the idea that industry-specific negative demand shocks reduce T but also negatively affect local manufacturing employment. As discussed in section 2.4, the literature has addressed this problem by using other high-wage countries' changing industry-specific imports (exports) from (to) China as an instrument (Z). The identifying assumption is that industry-specific demand trajectories are country-specific.[39] This is simply the identifying assumption $Z \perp\!\!\!\perp V_T$ that the literature following Autor et al. (2013) has invoked, as did we to obtain our TSLS estimates in section 3. For our purposes, we need to assume this identifying assumption is valid, and refer the reader to section 2.4 and the papers cited therein for further discussion. While other countries' import trajectories (Z) are thus assumed to be statistically independent of V_T, the same may not be true once we condition on T. For example, observing that imports of Chinese consumer electronics into Australia, Japan, and New Zealand (Z) begin to rise *relative to* Germany's imports from China in that product category may exactly proxy lower German demand over this time period.

[39] For this reason, we follow Dauth et al. (2014) in not using other Eurozone countries, instead following their recommendation of using Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the U.K.

4.3 The Restricted Mediation Model Under Linearity

In this section we discuss the identification and estimation of causal parameters of the restricted mediation model (Table 9) under the assumption that functions f_T, f_U, f_M, f_Y are linear. We show that, under linearity, mediation effects can be evaluated using the well-known method of Two-stage Least Squares (TSLS). Let the model equations of Panel B in Table 9 be defined as:

$$T = \xi_Z \cdot Z + \xi_V \cdot V_T + \epsilon_T, \tag{20}$$
$$U = \zeta_T \cdot T + \epsilon_U, \tag{21}$$
$$M = \varphi_T \cdot T + \varphi_U \cdot U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{22}$$
$$Y = \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V_Y + \epsilon_Y, \tag{23}$$

where $\xi_Z, \xi_V, \zeta_T, \varphi_T, \varphi_U, \delta_Y, \delta_Y, \delta_T, \beta_T, \beta_M, \beta_U, \beta_V$ are scalar coefficients, $\epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y$ are unobserved error-terms, Z, T, M, Y are observed, V_T, V_Y, U are unobserved and variables Z, V_T, V_M, $\epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y$ are mutually independent variables. We also assume that all variables have zero mean and that U, V_T, V_Y have unit variance without loss of generality (see Online Appendix H for detailed description).

Under Model (20)–(23), the counterfactual variables $M(t), Y(t), Y(m,t)$ are given by:[40]

$$M(t) = \Lambda_T^M \cdot t + \epsilon_{M(t)} \quad \text{where } \Lambda_T^M = \big(\varphi_T + \varphi_U \zeta_T\big) \tag{24}$$
$$\text{and } \epsilon_{M(t)} = \varphi_U \cdot \epsilon_U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M.$$
$$Y(t) = \Lambda_T^Y \cdot t + \epsilon_{Y(t)} \quad \text{where } \Lambda_T^Y = \Big(\beta_T + \beta_M \big(\varphi_T + \varphi_U \zeta_T\big) + \beta_U \zeta_T\Big) \tag{25}$$
$$\text{and } \epsilon_{Y(t)} = \beta_M \delta_T \cdot V_T + \beta_M \cdot \epsilon_M + \beta_U \cdot \epsilon_U + (\beta_V + \beta_M \delta_Y) \cdot V_Y + \epsilon_Y.$$
$$Y(m,t) = \Pi_M^Y \cdot m + \Pi_T^Y \cdot t + \epsilon_{Y(m,t)} \quad \text{where } \Pi_T^Y = \Big(\beta_T + \beta_U \zeta_T\Big), \quad \Pi_M^Y = \beta_M \tag{26}$$
$$\text{and } \epsilon_{Y(m,t)} = \beta_U \cdot \epsilon_U + \beta_V \cdot V_Y + \epsilon_Y.$$

We are interested in identifying four causal parameters $\Lambda_T^Y, \Lambda_T^M, \Pi_T^Y$ and Π_M^Y. Parameter Λ_T^Y in (24) denotes the total effect of T on Y, while Λ_T^M in (25) denotes the total effect of T on M. According to Equation (12), Parameter Π_T^Y in (26) stands for the *direct effect* of T on Y, while the *indirect effect* of T on Y is given by the product $\Pi_M^Y \cdot \Lambda_T^M$.[41] We prove the identification of each causal parameters in Online Appendix H.2.

Property 1 of **T-1** suggests that the TSLS of T on Y using Z as instrument estimates Λ_T^Y.[42]

[40] See Online Appendix H for a detailed derivation.

[41] See Online Appendix H.1 for further discussion on these causal parameters.

[42] See Online Appendix I for a detailed discussion on the estimation of the causal parameters of Model (20)–(23).

Equation (27) describe the TSLS while Equation (28) corroborates the estimated parameter.

$$\text{First Stage: } T = \Gamma^T + \Gamma^T_Z \cdot Z + \epsilon^T; \quad \text{Second Stage: } Y = \Gamma^Y + \Gamma^Y_T \cdot T + \epsilon^Y. \tag{27}$$

$$\text{plim}(\hat{\Gamma}^Y_T) = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,T)} = \frac{\big(\beta_T + \beta_M\big(\varphi_T + \varphi_U\zeta_T\big) + \beta_U\zeta_T\big)\xi_Z}{\xi_Z} = \Lambda^Y_T. \tag{28}$$

By the same token, Property 2 of **T-1** suggests that the TSLS of T on M using Z as instrument estimates Λ^M_T. Equation (29) describe this TSLS and Equation (30) confirms the estimated parameter.[43]

$$\text{First Stage: } T = \Gamma^T + \Gamma^T_Z \cdot Z + \epsilon^T; \quad \text{Second Stage: } M = \Gamma^M + \Gamma^M_T \cdot T + \epsilon^M. \tag{29}$$

$$\text{plim}(\hat{\Gamma}^M_T) = \frac{\text{cov}(Z,M)}{\text{cov}(Z,T)} = \frac{\big(\varphi_T + \varphi_U\zeta_T\big) \cdot \xi_Z}{\xi_Z} = \Lambda^M_T. \tag{30}$$

Property 3 of **T-1** suggests that the causal effect of M on Y can be estimated by the TSLS of M on Y conditioned on T that uses Z as instrument. Equations (31)–(32) specify this TSLS regression. Equations (33)–(33) show that the causal effect of M on Y, that is, Π^Y_M, is estimated by the coefficient of variable M in the second stage (32). Equations (34)–(35) explain that the causal parameter Π^Y_T is estimated by the the coefficient of variable T in the second stage (32). [44]

$$\text{First Stage: } M = \Gamma^{M|T} + \Gamma^{M|T}_Z \cdot Z + \Gamma^{M|T}_T \cdot T + \epsilon^{M|T}; \tag{31}$$

$$\text{Second Stage: } Y = \Gamma^{Y|T} + \Gamma^{Y|T}_M \cdot M + \Gamma^{Y|T}_T \cdot T + \epsilon^{Y|T}. \tag{32}$$

$$\text{plim}(\hat{\Gamma}^{Y|T}_M) = \frac{\text{cov}(T,Z) \cdot \text{cov}(T,Y) - \text{cov}(T,T) \cdot \text{cov}(Z,Y)}{\text{cov}(M,T) \cdot \text{cov}(T,Z) - \text{cov}(T,T) \cdot \text{cov}(Z,M)} \tag{33}$$

$$= \frac{\text{cov}(V_T,V_T) \cdot \beta_M \cdot \delta_T \cdot \xi_Z}{\text{cov}(V_T,V_T) \cdot \delta_T \cdot \xi_Z} = \beta_M = \Pi^Y_M; \tag{34}$$

$$\text{plim}(\hat{\Gamma}^{Y|T}_T) = \frac{-\big(\text{cov}(M,Z) \cdot \text{cov}(T,Y) - \text{cov}(M,T) \cdot \text{cov}(Z,Y)\big)}{\text{cov}(M,T) \cdot \text{cov}(T,Z) - \text{cov}(T,T) \cdot \text{cov}(Z,M)} \tag{35}$$

$$= \frac{\text{cov}(V_T,V_T) \cdot \delta_T \cdot \xi_Z \cdot \big(\beta_T + \beta_U \cdot \zeta_T\big)}{\text{cov}(V_T,V_T) \cdot \delta_T \cdot \xi_Z} = \big(\beta_T + \beta_U \cdot \zeta_T\big) = \Pi^Y_T. \tag{36}$$

In Section (4.1) we explain that the exclusion restriction $Z \perp\!\!\!\perp Y(m)$ does not hold for the mediation model of Table 9. Thus we should expect that the TSLS of M on Y that uses Z as instrument does not render a causal parameter. Equation (37) reports this TSLS estimate, which

[43] Online Appendix I describes the estimation of this causal parameter in greater detail.

[44] See Online Appendix H.2 for a detailed derivation of Equations (33)–(36). See Online Appendix I for a description of the TSLS estimation.

does not have a clear causal interpretation.

$$\frac{\mathrm{cov}(Z,Y)}{\mathrm{cov}(Z,M)} = \frac{\big(\varphi_T + \varphi_U\zeta_T\big)\cdot\xi_Z}{\Big(\beta_T + \beta_M\big(\varphi_T + \varphi_U\zeta_T\big) + \beta_U\zeta_T\Big)\cdot\xi_Z}, \tag{37}$$

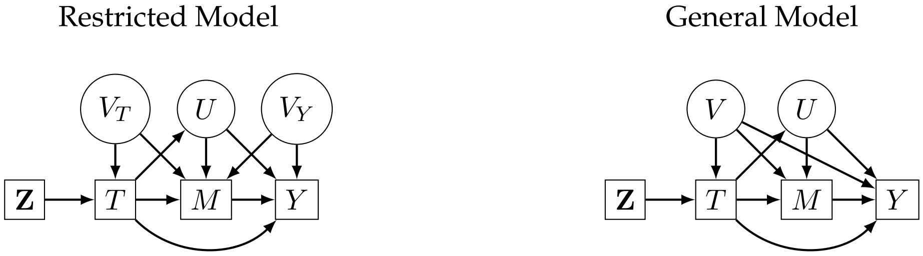
4.4 A Model Specification Test

The restricted mediation model is build by imposing restrictions on the causal links of the unobserved confounder V of the general mediation model. In this section we discuss a simple method to test these causal restrictions.

Table 10 displays the restricted and general models as DAGs (Panel A) and the linear equations that subsume the causal relations of each model (Panel B). Our null hypothesis is that the data generating process conforms to the restricted mediation model. Our alternative hypothesis is that the general mediation model holds. Effectively, we test if the causal assumption that splits the confounder V of the general model into V_T and V_Y in the restricted model is empirically sound. To do so, we evoke the linearity. Our aim is not to test linearity itself, but rather use linearity to do inference on the causal assumptions of restricted model.

Table 10: Restricted and General Mediation Model with One Instrumental Variable

A. DAG Representation



Restricted Model

General Model

B. Linear Equations

$T = \boldsymbol{\xi}_Z \cdot \mathbf{Z} + \xi_V \cdot V_T + \epsilon_T$	$T = \boldsymbol{\xi}_Z \cdot \mathbf{Z} + \xi_V \cdot V + \epsilon_T$
$U = \zeta_T \cdot T + \epsilon_U$	$U = \zeta_T \cdot T + \epsilon_U$
$M = \varphi_T \cdot T + \varphi_U \cdot U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M$	$M = \varphi_T \cdot T + \varphi_U \cdot U + \delta \cdot V + \epsilon_M$
$Y = \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V_Y + \epsilon_Y$	$Y = \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V + \epsilon_Y$

Panel A presents the Directed Acyclic Graphs (DAG) of the restricted and general models. Panel B presents the equations that subsume the causal relations described in each model under the assumption of linearity.

The identification of coefficients in linear models depends on the equations governing the covariance structure of observed data. Our test explores the differences on these equations between the restricted and the general models. Online Appendix K shows that Equalities (**??**)–(**??**) must hold for the restricted model.

$$\mathrm{cov}(\mathbf{Z},Y) - \mathrm{cov}(\mathbf{Z},M)\cdot\beta_M - \mathrm{cov}(\mathbf{Z},T)\cdot\widetilde{\beta}_T = 0 \tag{38}$$

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,M)\cdot\beta_M - \mathrm{cov}(T,T)\cdot\widetilde{\beta}_T = 0, \tag{39}$$

where $\widetilde{\beta}_T = \varphi_T + \varphi_U\zeta_T$. Equalities (**??**)–(**??**) enable the identification of parameters $\widetilde{\beta}_T$ and β_M. If the instrument Z consists of a single variable, then parameters $\widetilde{\beta}_T$ and β_M are exactly identified and our model specification test does not apply (see Online Appendix K.1). If Z consists of two (or more) variables, then parameters $\widetilde{\beta}_T$ and β_M are over-identified and Equation (**??**) constitute an over-identification restriction (see Online Appendix K.2).[45] The respective equalities for the case of the general mediation model are derived in Online Appendix K.3 and are given by:

$$\mathrm{cov}(\mathbf{Z},Y) - \mathrm{cov}(\mathbf{Z},M)\cdot\beta_M - \mathrm{cov}(\mathbf{Z},T)\cdot\widetilde{\beta}_T = 0 \tag{40}$$

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,M)\cdot\beta_M - \mathrm{cov}(T,T)\cdot\widetilde{\beta}_T = \beta_V\cdot\xi_V. \tag{41}$$

Our inference explores the fact that the over-identification restriction (**??**) is equal to zero for restricted model while restriction (**??**) differs from zero in the case of the general model. Recall that model variables have mean zero w.l.o.g., thus we can express Equalities (**??**)–(**??**) as the following moment conditions:

$$E(\mathbf{Z}\cdot(Y-\beta_M\cdot M-\widetilde{\beta}_T\cdot T))=0 \quad \text{for} \quad \mathrm{cov}(\mathbf{Z},Y) - \mathrm{cov}(\mathbf{Z},M)\cdot\beta_M - \mathrm{cov}(\mathbf{Z},T)\cdot\widetilde{\beta}_T = 0 \tag{42}$$

$$\text{and } E(T\cdot(Y-\beta_M\cdot M-\widetilde{\beta}_T\cdot T))=0 \quad \text{for} \quad \mathrm{cov}(T,Y) - \mathrm{cov}(T,T)\cdot\widetilde{\beta}_T - \mathrm{cov}(T,M)\cdot\beta_M = 0 \tag{43}$$

Moment Conditions (**??**)–(**??**) can be used to estimate parameters $\beta_M, \widetilde{\beta}_T$ via the Generalized Method of Moments (GMM) of Hansen (1982).

Our inference is based on a simple rationale. If multiple instruments exists, then parameters $\beta_M, \widetilde{\beta}_T$ can be consistently estimated using Moment Condition (**??**). This condition is associated to Equations (**??**),(**??**) and thereby holds for both restricted and general models. On the other hand, Moment Condition (**??**) is valid only for the restricted model. Thus a model specification test consists on verifying if the model estimates comply with Moment Condition (**??**). Large ab-

[45] Our test bares some some similarities with the the Sargan-Hansen test that exploits model over-identifying restrictions to do inference on model coefficients.

solute values of Moment Condition (**??**) constitute statistical evidence contrary to the restricted mediation model and the null hypothesis is rejected.

We implement this test by evaluating two sets of GMM estimators for parameters $\beta_M, \widetilde{\beta}_T$. The first estimator is based only on Moment Condition (**??**). The second estimators relies on both moment conditions (**??**) and (**??**). Large absolute differences between these estimates provide statistical evidence gainst the null hypothesis that data is generated by the restrictive model. Thus we then perform a Wald test on the null hypothesis that the two GMM estimates are equal. See Online Appendix K.4 for a description of the inference procedure.

5 Mediation Analysis

In this section, we implement the mediation model outlined in section 4 in order to estimate the *indirect effect* of trade exposure on voting that works through labor markets. The extent to which trade exposure polarized voters because it caused labor market turmoil is identified by a comparison of the *indirect effect* with the *total effect* of trade exposure on voting.

For voting Y, we focus on the vote share of extreme right parties, i.e. the only significant voting response to trade observed in the data. The *total effect* $\hat{\Gamma}_T^Y$ is reported in table 3. For the mediator M we focus on the second principal component of the six labor market outcomes, i.e. the principal component causally affected by trade. The effect of T on M $\hat{\Gamma}_T^M$ is reported in table 6. To obtain the *indirect effect* we need to additionally estimate $\hat{\Gamma}_M^{Y|T}$, the effect of M on Y. As shown in section 4.3, we estimate $\hat{\Gamma}_M^{Y|T}$ in the following second stage equation:

$$Y_{it} = \Gamma_T^{Y|T} \cdot T_{it} + \Gamma_M^{Y|T} \cdot M_{it} + \Gamma_X^{Y|T} \cdot X_{it} + \epsilon_{it}^{Y|T}. \tag{44}$$

The First Stage now takes the form

$$M_{it} = \Gamma_{IM}^{M|T} \cdot Z_{it}^{IM} + \Gamma_{EX}^{M|T} \cdot Z_{it}^{EX} + \Gamma_T^{M|T} \cdot T_{it} + \Gamma_X^{M|T} \cdot X_{it} + \epsilon_{it}^{M|T}. \tag{45}$$

(51) differs from the second stage equation (3) in that we are interested in the causal effect of the mediator M on outcome Y (instead of T on Y or T on M). (52) differs from the first stage equation (4) in that trade exposure T is now included. (See section 4.3.) The *indirect effect* is given

Table 11: First Stage Relationship (Equation (52))

	(1)	(2)	(3)	(4)	(5)	(6)
Z^{IM}_{it}	0.002 (0.088)	-0.022 (-0.963)	-0.028 (-1.319)	-0.036* (-1.777)	-0.034* (-1.787)	-0.034* (-1.787)
Z^{EX}_{it}	0.053** (2.216)	0.056** (2.466)	0.069*** (3.238)	0.071*** (3.477)	0.070*** (3.442)	0.070*** (3.442)
$\mathbf{T}_{it}$	-0.057 (-1.187)	-0.075 (-2.392)	-0.062 (-1.410)	-0.091*** (-2.942)	-0.075 (-1.407)	-0.075 (-1.407)
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: The table presents the first stage results of estimating equation (52). For brevity, only coefficients for $Z^{IM}_{it}, Z^{EX}_{it}, T_{it}$ are reported. The table's structure and data-observations are identical to tables 3 and 6. *t-statistics* reported in brackets. Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1.

by multiplying $\Gamma_M^{Y|T}$ from with $\hat{\Gamma}_T^M$. The *direct effect* is not identified as a causal parameter in (51). Instead it is given by $\hat{\Gamma}_T^Y - \Gamma_M^{Y|T} \times \hat{\Gamma}_T^M$.

Table 11 reports the results of estimating the First Stage equation (52). As discussed in section 4.2, the fact that the TSLS estimates $\hat{\Gamma}_T^M$ in table 6 are larger than the OLS is suggestive of domestic industry-specific demand conditions as a source of confounding bias V_T: German industries that experience negative demand shocks will see fewer imports and less employment. The discussion in section 4.2 implies that $Z|T$ can serve as a proxy for V_T, assuming it has enough explanatory power. This suggests that, conditional on Germany's imports, other countries' imports ($Z|T$) should "cause" additional turmoil in German labor markets. Indeed, this is what we find. Other similar countries' imports from low-wage manufacturers reduce the second principal component of German labor markets while exports increase it. As before we use two designated instruments Z^{IM}_{it}, Z^{EX}_{it}. Two designated instruments (for T) are necessary for the model specification test in 4.4.

The top-panel of table 12 reports the TSLS estimates of (51). Estimate $\hat{\Gamma}_M^{Y|T}$ gives the effect of labor market disturbances M_{it} on extreme right party vote shares Y_{it}. M_{it} is the second principal component of six observed labor market outcomes, standardized to have a mean of zero an a standard-deviation of one. Decreases of M_{it} indicate worsening labor markets specifically for manufacturing workers, as shown in table 5. $\hat{\Gamma}_M^{Y|T}$ suggests that a one-standard deviation decrease in the principal component increases the extreme right's vote share by 0.492 (column 5). The effect of trade-induced labor market disturbances on voting, i.e. the *indirect* or *mediated* effect of T on Y, can be derived by multiplying $\hat{\Gamma}_M^{Y|T}$ by $\hat{\Gamma}_T^M$. This is equivalent to combining table 6

Table 12: Effect of M_{it} on Y_{it} (Equation (51))

	(1) Baseline IV	(2) + Structure IV	(3) + Industry IV	(4) + Voting IV	(5) +Socio IV	(6) Standard. IV
Second Stage:						
$\mathbf{M}_{\text{it}}$	-0.473** (-2.038)	-0.585** (-2.144)	-0.534** (-2.406)	-0.504*** (-3.138)	-0.492*** (-2.900)	-0.290*** (-2.900)
$\mathbf{T}_{\text{it}}$	-0.057 (-1.187)	-0.075 (-1.413)	-0.062 (-1.410)	-0.091*** (-2.942)	-0.086*** (-2.632)	-0.043*** (-2.632)
% Effect Mediated by $\mathbf{M}_{\text{it}}$	106	178	155	190	178	150
Specification Test						
$\widehat{\beta_{M,2}}$	-0.294 [0.199]	-0.608 [0.176]	-0.527 [0.152]	-0.308 [0.267]	-0.263 [0.334]	-0.155 [0.334]
$\widehat{\beta_{M,1}} - \widehat{\beta_{M,2}}$	-0.179 [0.584]	0.023 [0.965]	-0.007 [0.988]	-0.196 [0.541]	-0.229 [0.476]	-0.135 [0.4762]
$\widehat{\beta_{T,2}}$	0.040 [0.465]	-0.084 [0.502]	-0.059 [0.597]	-0.014 [0.866]	0.005 [0.956]	0.002 [0.956]
$\widehat{\beta_{T,1}} - \widehat{\beta_{T,2}}$	-0.096 [0.183]	0.009 [0.949]	-0.003 [0.982]	-0.077 [0.385]	-0.090 [0.308]	-0.045 [0.308]

Notes: (*a*) The table presents the first stage results of estimating equation (52). The table's structure and data-observations are identical to tables 3 and 6. Only the coefficient on M_{it} has a causal interpretation. (*b*) At the bottom of the Second Stage panel we report the ratio of the *indirect effect* and the *total effect*. In column 5 the indirect effect combines estimates from table 12 and table 6: $0.1584 = -0.492 \cdot -0.322$. The *total effect* is estimated as 0.089 in table 3. (*c*) The Specification Test panel reports estimates from the General Mediation Model and compares them to our results from the Restricted Mediation Model in the he top panel, e.g. $-0.229 = -0.492 + 0.263$. This test fails to reject the validity of our identifying assumptions: The difference in the estimated coefficients from the two models is not significant. (*d*) Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1. The Second Stage panel reports *t-statistics* in brackets. The Specification Test panel reports *p-values* in brackets.

column 5 with table 12 column 5. The implied magnitude of the indirect effect $\hat{\Gamma}_M^{Y|T} \cdot \hat{\Gamma}_T^M$ is 0.1584 $(-0.492 \cdot -0.322)$. A one-standard deviation increase in trade exposure T_{it} (1,350 €) induces labor market disturbances which in turn increase the extreme-right vote share by 0.213 $(0.1584 \cdot 1.35)$ percentage points, i.e. roughly 50 percent of the average per-decade increase of 0.43 percentage points during the 22 years we study. The percentage of the total populist effect that is explained by labor markets is $(\hat{\Gamma}_M^{Y|T} \cdot \hat{\Gamma}_T^M)/\hat{\Gamma}_T^Y$. It ranges from 155% to 190% across columns 2–5. This means that labor market disturbances easily explain the entirety of the effect of trade on voting. Our mediation analysis thus shows that effect of trade exposure on voters' support for populist parties is entirely driven by a large polarizing effect that runs through labor market disturbances.

The 'direct' effect of trade on voting $\hat{\Gamma}_T^{Y|T}$ is identified as the residual $\hat{\Gamma}_T^Y - (\hat{\Gamma}_M^{Y|T} \cdot \hat{\Gamma}_T^M)$ i.e. $0.089 - (-0.492 \cdot -0.322) = -0.0694$. Other suggested channels that may cause a voter response to trade– e.g. goods price reductions, expansions in product variety, government programs, or anticipation effects–appear to have been politically moderating when considered together as a bundle.[46]

The bottom-panel of table 12 (*Specification Test*) presents the model specification test that compares estimates $\widehat{\beta_{M,1}}$ from (197) versus $\widehat{\beta_{M,2}}$ from (199) and , $\widehat{\widetilde{\beta}_{T,1}}$ from (197) versus $\widehat{\widetilde{\beta}_{T,2}}$ from (199). Our null hypothesis is that the causal assumptions of the restricted model hold. Under the null, $\widehat{\beta_{M,1}}$ and $\widehat{\beta_{M,2}}$ are consistent estimators of β_M. Thereby we test if the difference $\widehat{\beta_{M,1}} - \widehat{\beta_{M,2}}$ is statistically significant using a Wald test that is based on the χ^2-statistic associated with this difference. Estimates are generated by a two-step GMM estimator whose errors are clustered by region. We use the same set of conditioning variables described in the main paper. We perform the same procedure to the difference $\widehat{\widetilde{\beta}_{T,1}} - \widehat{\widetilde{\beta}_{T,2}}$. We do not reject the null hypothesis that the causal assumptions of the restricted model hold in any of these single hypothesis tests.

6 Discussion & Conclusion

Trade liberalization creates distributional frictions between its winners and losers. A substantial body of recent evidence suggests that in high-wage manufacturing countries like Germany and the U.S., the losers are primarily manufacturing workers. At the same time trade liberalization appears to have lead to have been a boon to populist and protectionist politicians and parties. In this paper we ask to what extent the labor market consequences of trade liberalization are the reason for this rise in political populism.

Using German data for 1987–2009, we indeed find that higher import competition has led to increases in the vote-share of populist parties, in Germany's case this has meant the extreme right. As well, we find that import competition has had detrimental effects on manufacturing workers.

The focus of our paper is to understand to what extent the labor market disturbances caused by trade liberalization are the reason that trade liberalization has led to an increase in populist voting. To answer this question, which When we try to understand the underlying mechanisms, we face a common empirical problem: Even though we can causally identify the total effect of the treatment (T_{it}) on voting as the final outcome (Y_{it}) *and* on labor markets as a proposed mechanism (M_{it}), we

[46] Our analysis can not shed further light on unpacking the effects of these other suggested channels. To do so, we would need to (a) observe them and (b) have separate designated instruments for each one.

cannot easily identify how much of the total effect works through the observed mechanism.

We develop a new methodology that allows us to perform the required *mediation analysis* in an IV setting like ours. We find that the populist voter response to trade liberalization is entirely explained by the labor market disturbances it has caused. We thus clearly identify labor markets disturbances as the most important reason for the political backlash against free trade.

Our empirical framework may be useful in a broad range of empirical applications studying causal mechanisms in IV settings.

References

Altonji, J. G. and R. L. Matzkin (2005, July). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Art, D. (2007). Reacting to the radical right lessons from germany and austria. *Party Politics 13*(3), 331–349.

Arzheimer, K. (2009). Contextual Factors and the Extreme Right Vote in Western Europe, 1980-2002. *American Journal of Political Science 53*(2), 259–275.

Autor, D., D. Dorn, and G. Hanson (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review 103*(6), 2121–68.

Autor, D., D. Dorn, G. Hanson, and K. Majlesi (2016). Importing political polarization? the electoral consequences of rising trade exposure. *NBER Working Paper*.

Bade, F.-J., C.-F. Laaser, and R. Soltwedel (2003). Urban specialization in the internet ageempirical findings for germany, processed. *Kiel Institute for World Economics*.

Bagues, M. and B. Esteve-Volart (2014). Politicians' Luck of the Draw: Evidence from the Spanish Christmas Lottery. *Accepted at Journal of Political Economy*.

Bender, S., A. Haas, and C. Klose (2000). Iab employment subsample 1975-1995 opportunities for analysis provided by the anonymised subsample. *IZA Discussion Paper 117*.

Blundell, R. and J. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In L. P. H. M. Dewatripont and S. J. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Volume 2. Cambridge, UK: Cambridge University Press.

Blundell, R. and J. Powell (2004, July). Endogeneity in semiparametric binary response models. *Review of Economic Studies 71*(3), 655–679.

Brunner, E., S. L. Ross, and E. Washington (2011). Economics and policy preferences: causal evidence of the impact of economic conditions on support for redistribution and other ballot proposals. *Review of Economics and Statistics 93*(3), 888–906.

Charles, K. K. and M. J. Stephens (2013). Employment, wages, and voter turnout. *American Economic Journal: Applied Economics 5*(4), 111–143.

Che, Y., Y. Lu, J. R. Pierce, P. K. Schott, and Z. Tao (2016). Does trade liberalization with china influence us elections? Technical report, National Bureau of Economic Research.

Dauth, W., S. Findeisen, and J. Suedekum (2014). The Rise of the East and the Far East: German Labor Markets and Trade Integration. *Journal of European Economic Association 12*(6), 1643–1675.

Dippel, C., R. Gold, and S. Heblich (2015). Globalization and its (dis-) content: Trade shocks and voting behavior. *NBER Working Paper* (w21812).

Dix-Carneiro, R., R. R. Soares, and G. Ulyssea (2017). Economic shocks and crime: Evidence from the brazilian trade liberalization. Technical report, National Bureau of Economic Research.

Duranton, G. and D. Puga (2005). From sectoral to functional urban specialisation. *Journal of Urban Economics 57*(2), 343–370.

Dustmann, C., B. Fitzenberger, U. Schönberg, and A. Spitz-Oener (2014). From sick man of europe to economic superstar: Germany's resurgent economy. *The Journal of Economic Perspectives 28*(1), 167–188.

Falck, O., R. Gold, and S. Heblich (2014). E-lections: Voting Behavior and the Internet. *American Economic Review 104*(7), 2238–65.

Falck, O., S. Heblich, and A. Otto (2013). Agglomerationsvorteile in der wissensgesellschaft: Empirische evidenz für deutsche gemeinden. *ifo Schnelldienst 66*(3), 17–21.

Falk, A., A. Kuhn, and J. Zweimüller (2011). Unemployment and Right-wing Extremist Crime. *The Scandinavian Journal of Economics 113*(2), 260–285.

Feigenbaum, J. J. and A. B. Hall (2015). How legislators respond to localized economic shocks: Evidence from chinese import competition. *Journal of Politics 77*(4), 1012–30.

Frank, T. (March 7th 2016). Millions of ordinary americans support donald trump. here's why. *The Guardian*.

Frölich, M. and M. Huber (2014). Direct and indirect treatment effects: causal chains and mediation analysis with instrumental variables. *IZA Working Paper* (8280).

Giuliano, P. and A. Spilimbergo (2014). Growing up in a recession. *The Review of Economic Studies 81*(2), 787–817.

Grumke, T. (2012). *The Extreme Right in Europe*. Vandenhoeck & Ruprecht.

GSOEP (2007). The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements. Technical Report 1.

Hafeneger, B. and S. Schönfelder (2007). *Politische Strategien gegen die extreme Rechte in Parlamenten. Folgen für kommunale Politik und lokale Demokratie*. Friedrich-Ebert-Stiftung: Berlin.

Hagan, J., H. Merkens, and K. Boehnke (1995). Delinquency and Disdain: Social Capital and the Control of Right-Wing Extremism Among East and West Berlin Youth. *American Journal of Sociology 100*(4), 1028–1052.

Hansen, L. P. (1982, July). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Heckman, J. and R. Pinto (2017). Unordered monotonicity. *Forthcoming Econometrica*.

Heckman, J. J. (2008). The principles underlying evaluation estimators with an application to matching. *Annales d'Economie et de Statistiques 91–92*, 9–73.

Heckman, J. J. and R. Pinto (2015a). Causal analysis after Haavelmo. *Econometric Theory 31*(1), 115–151.

Heckman, J. J. and R. Pinto (2015b). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric reviews 34*(1-2), 6–31.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Hiscox, M. J. (2002). Commerce, coalitions, and factor mobility: Evidence from congressional votes on trade legislation. *American Political Science Review 96*(3), 593–608.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011a). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*, 765–789.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011b). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*(4), 765–789.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Imbens, G. W. (2004, February). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*(1), 4 – 29.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and W. K. Newey (2007). Identification and estimation of triangular simultaneous equations models without additivity. Unpublished manuscript, Harvard University and MIT.

Jensen, J. B., D. P. Quinn, and S. Weymouth (2016). Winners and losers in international trade: The effects on us presidential voting. Technical report, National Bureau of Economic Research.

Krueger, A. B. and J.-S. Pischke (1997). A Statistical Analysis of Crime Against Foreigners in Unified Germany. *Journal of Human Resources 32*(1), 182–209.

Krugman, P. R. (2008). Trade and Wages, Reconsidered. *Brookings Papers on Economic Activity 2008*(1), 103–154.

Lubbers, M. and P. Scheepers (2001). *European Sociological Review 17*(4), 431–449.

Malgouyres, C. (2014). The impact of exposure to low-wage country competition on votes for the far-right: Evidence from french presidential elections. *working paper*.

Malgouyres, C. (2017). The impact of chinese import competition on the local structure of employment and wages: Evidence from france. *Journal of Regional Science 57*(3), 411–441.

Mocan, N. H. and C. Raschke (2014). Economic Well-being and Anti-Semitic, Xenophobic, and Racist Attitudes in Germany. *National Bureau of Economic Research Working Paper 20059*.

Mudde, C. (2000). *The Ideology of the Extreme Right*. Manchester University Press.

Mughan, A., C. Bean, and I. McAllister (2003). Economic globalization, job insecurity and the populist reaction. *Electoral Studies 22*(4), 617–633.

Mughan, A. and D. Lacy (2002). Economic Performance, Job Insecurity and Electoral Choice. *British Journal of Political Science 32*(3), 513–533.

New York Times (2009). Ancient citys nazi past seeps out after stabbing. *February 11th*.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods, Special Section: Naturally Occurring Section on Causation Topics in Psychological Methods 19*, 459–481.

Petersen, M. L., S. E. Sinisi, and M. J. Van der Laan (2006). Estimation of direct causal effects. *Epidemiology 17*, 276–284.

Pierce, J. R. and P. K. Schott (2016). The surprisingly swift decline of us manufacturing employment. *American Economic Review 106*(7), 1632–62.

Pinto, R. (2015). Selection bias in a controlled experiment: The case of Moving to Opportunity. Unpublished Ph.D. Thesis, University of Chicago, Department of Economics.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In N. L. P. J. Green, Hjort and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, MR2082403, pp. 70–81. Oxford: Oxford University Press.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*(2), 143–155.

Rodrik, D. (1995). Political economy of trade policy. *Handbook of international economics 3*(3), 1457–1494.

Rogowski, R. (1987). Political cleavages and changing exposure to trade. *American Political Science Review 81*(4), 1121–1137.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes (with discussion). *Scandinavian Journal of Statistics 31*, 161–170.

Scheve, K. F. and M. J. Slaughter (2001). What Determines Individual Trade-Policy Preferences? *Journal of International Economics 54*(2), 267–292.

Sommer, B. (2008). Anti-capitalism in the name of ethno-nationalism: ideological shifts on the german extreme right. *Patterns of Prejudice 42*(3), 305–316.

Stöss, R. (2010). Rechtsextremismus im Wandel. Technical report, Friedrich Ebert Stiftung.

The Economist (July 30th 2016). The new political divide.

Voigtländer, N. and H.-J. Voth (2015). Taught to Hate: Nazi Indoctrination and Anti-Semitic Beliefs in Germany. *Proceedings of the National Academy of Sciences Forthcoming*.

Vytlacil, E. J. (2006, August). Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *Review of Economics and Statistics 88*(3), 578–581.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT press.

Appendix A Identification of T on Y with Gravity Residuals

Table 13: Gravity Results for Effect of T_{it} on Voting

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	+ Structure	+ Industry	+ Voting	+Socio	Standard.
	Gravity	Gravity	Gravity	Gravity	Gravity	Gravity
Δ Turnout	0.000**	0.000*	0.000**	0.000**	0.000*	0.002*
	(2.143)	(1.774)	(1.980)	(1.966)	(1.706)	(1.706)
Established Parties:						
Δ Vote Share CDU/CSU	0.006	0.003	0.003	0.001	0.001	0.000
	(0.873)	(0.392)	(0.342)	(0.186)	(0.089)	(0.089)
Δ Vote Share SPD	-0.008	-0.004	-0.005	0.004	0.004	0.000
	(-1.197)	(-0.674)	(-0.720)	(0.732)	(0.668)	(0.668)
Δ Vote Share FDP	0.001	0.006	0.004	0.007*	0.007*	0.001*
	(0.204)	(1.365)	(1.064)	(1.840)	(1.933)	(1.933)
Δ Vote Share Green Party	0.006**	0.000	0.001	0.000	0.000	0.000
	(2.021)	(0.071)	(0.408)	(0.047)	(0.012)	(0.012)
Non-established Parties						
Δ Vote Share Extreme-Right Parties	0.004*	0.006**	0.006**	0.003	0.003*	0.002*
	(1.855)	(2.430)	(2.276)	(1.575)	(1.779)	(1.779)
Δ Vote Share Far-Left Parties	-0.011*	-0.012*	-0.012*	-0.014**	-0.013**	-0.003**
	(-1.814)	(-1.884)	(-1.894)	(-2.382)	(-2.177)	(-2.177)
Δ Vote Share Other Small Parties	0.002	0.002	0.003	-0.001	-0.001	-0.001
	(0.755)	(0.696)	(1.141)	(-0.710)	(-0.766)	(-0.766)
Period-by-region F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: (*a*) Each cell reports results from a separate regression. The data is a stacked panel of first-differences at the *Landkreis* level. Each regression has 730 observations, i.e. 322 *Landkreise* in West Germany, observed in 1987–1998 and 1998–2009, and 86 *Landkreise* in East Germany, observed only in 1998–2009. We drop three city-states (Hamburg, Bremen, and Berlin in the East). (*b*) All specifications include region-by-period fixed effects. Column 1 controls only for start-of-period manufacturing. Column 2 adds controls for the structure of the workforce (share female, foreign, and high-skilled). Column 3 adds controls for dominant industries (employment share of the largest industry, in automobiles, and chemicals). Column 4 adds start-of-period voting controls. Column 5 adds socioeconomic controls at the start of the period (population share of unemployed individuals, and individuals aged 65+). This is our preferred specification. Finally, Column 6 presents our preferred specification with standardized outcome variables to facilitate comparison. (*c*) All standard errors are clustered at the level of 96 commuting zones. All specifications include region-by-period fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

An alternative approach to the IV approach pursued in the paper is to estimate gravity equations, which exploit essentially the same source of exogenous variation. The endogeneity concern with increasing imports ΔIM_{Gjt} is that they reflect not only increasing competitiveness of Chinese and Eastern European ('CE') industries[47], but also German industry-specific demand changes.

The gravity approach to solving this problem is to compare changes in German industries' exports to other countries O in relation to Chinese and Eastern European exports to O. This comparison reflects changes in Chinese and Eastern European comparative advantage over Germany, and allows constructing an exogenous measure ΔIM_{Gjt}^{grav} to replace ΔIM_{Gjt}.[48]

Online Appendix D shows how to obtain the gravity-residuals ΔIM_{Gjt}^{grav} that replace ΔIM_{Gjt} the gravity-residuals ΔEX_{Gjt}^{grav} that replace ΔEX_{Gjt}. An exogenous measure for changes in in German industries' trade exposure can be obtained from netting out both effects such that $\Delta Trade_{Gjt}^{grav} = \Delta IM_{Gjt}^{grav} - \Delta EX_{Gjt}^{grav}$. Substituting $\Delta Trade_{Gjt}$ in equation (1) with $\Delta Trade_{Gjt}^{grav}$ provides an exogenous measure of regional trade exposure based on the gravity approach as

$$T_{it}^{grav} = \sum_j \frac{L_{ijt}}{L_{jt}} \frac{\Delta Trade_{Gjt}^{grav}}{L_{it}} \tag{46}$$

We now substitute T_{it} from our baseline regression (3) with T_{it}^{grav} directly. Otherwise, we run exactly the same specifications when estimating

$$Y_{it} = \Gamma_T^Y \cdot T_{it}^{grav} + \Gamma_X^Y \cdot X_{it} + \epsilon_{it}^Y \tag{47}$$

Results are reported in Table 13. Again, each cell reports results from a different regression. Rows specify different outcome variables, and columns refer to different regression specifications. Results are consistent with our main specifications reported in Table 3. Following the gravity approach, there is additional evidence for a positive effect of trade exposure on turnout. Moreover, the positive effect on the vote share of the market-liberal party FDP turns significant in our preferred specification in column 5. Additionally, the negative effect on far-left parties is significant in the gravity regressions. We do not want to over-interpret these results, but take them as an indication that trade exposure might indeed have more pronounced effects on voting behavior than our original identification strategy suggests. Most importantly, the positive effect of trade exposure on extreme-right party votes is confirmed by this alternative identification strategy.

[47] Competitiveness increases due to productivity increases, better market access, and decreasing relative trade cost.

[48] As before, we chose Belgium, France, Greece, Italy, Luxembourg, the Netherlands, Portugal, Spain, and the UK as ”other countries” O for our gravity regressions, to be comparable with Dauth et al. (2014).

Appendix B Proof of Theorem T-1

Proof P-1 *Our model stem from seven exogenous and statistically independent random variables: the unobserved five error terms $\epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y$, the observed instrument Z and the unobserved variables V_T, and V_Y. All remaining variables of the model can be expressed in term of these seven variables.*

We first show that non-independence relations $Z \not\perp\!\!\!\perp T$ and $Z \not\perp\!\!\!\perp M|T$ hold. To do so, it is useful to express the treatment T and the mediator variable M in terms of the model exogenous variables:

$$
\begin{align}
T &= f_T(Z, V_T, \epsilon_T), \tag{48}\\
\textit{and } M &= f_M(T, U, V_T, V_Y, \epsilon_M), \nonumber\\
\textit{but } U &= f_U(T, \epsilon_U), \nonumber\\
\textit{thus } &= f_M(T, f_U(T, \epsilon_U), V_T, V_Y, \epsilon_M). \tag{49}
\end{align}
$$

Equation (55) implies $Z \not\perp\!\!\!\perp T$. It remains to prove that $Z \not\perp\!\!\!\perp M|T$. To do so, it is useful to show that $Z \not\perp\!\!\!\perp V_T|T$. According to equation (55), conditioning on $T = t$ is equivalent to conditioning on the values of V_T, Z, ϵ_T such that $f_T(Z, V_T, \epsilon_T) = t$. This induces a correlation between Z and V_T and thereby $Z \not\perp\!\!\!\perp V_T|T$. While $Z \perp\!\!\!\perp (V_T, \epsilon_T)$ holds because (Z, V_T, ϵ_T) are statistically independent, $Z \perp\!\!\!\perp (V_T, \epsilon_T)|(T = t)$ does not.

The arguments of mediator M in (56) can be split into into two sets of variables: (T, V_T) and $(V_Y, \epsilon_U, \epsilon_M)$. Note that independence relation among exogenous variables $(Z, V_T, \epsilon_T) \perp\!\!\!\perp (V_Y, \epsilon_U, \epsilon_M)$ holds. But, according to (55), T is a function of (Z, V_T, ϵ_T). Thereby $T \perp\!\!\!\perp (V_Y, \epsilon_U, \epsilon_M)$ also holds. Nevertheless, the remaining arguments of M are T, V_T and $Z \not\perp\!\!\!\perp V_T|T$ implies that $Z \not\perp\!\!\!\perp g(T, V_T, V_Y, \epsilon_U, \epsilon_M)|T$ for any non-degenerate function $g(\cdot)$ of V_T and, in particular, $Z \not\perp\!\!\!\perp M|T$.

It remains to prove the three exclusion restrictions: (1) $Z \perp\!\!\!\perp Y(t)$; (2) $Z \perp\!\!\!\perp M(t)$; and (3) $Z \perp\!\!\!\perp Y(m)|T$. To so do, it is useful to express the counterfactuals $Y(t), M(t)$ and $Y(m)$ in terms of the seven *exogenous variables of the model:*

$$
\begin{aligned}
U(t) &= f_U(t, \epsilon_U), && (50)\\
M(t) &= f_M(t, U(t), V_T, V_Y, \epsilon_M)\\
&= f_M(t, f_U(t, \epsilon_U), V_T, V_Y, \epsilon_M) && (51)\\
Y(t) &= f_Y(t, M(t), U(t), V_Y, \epsilon_Y)\\
&= f_Y(t, f_M(t, f_U(t, \epsilon_U), V_T, V_Y, \epsilon_M), f_U(t, \epsilon_U), V_Y, \epsilon_Y) && (52)\\
Y(m) &= f_Y(T, m, U, V_Y, \epsilon_Y)\\
&= f_Y(T, m, f_U(T, \epsilon_U), V_Y, \epsilon_Y) && (53)
\end{aligned}
$$

According to equation (59), $Y(t)$ *is a function of exogenous variables* $\epsilon_U, V_T, V_Y, \epsilon_M, \epsilon_Y$, *thus:*

$$\Big(\epsilon_U, V_T, V_Y, \epsilon_M, \epsilon_Y\Big) \perp\!\!\!\perp Z \;\Rightarrow\; Y(t) \perp\!\!\!\perp Z.$$

According to (58), $M(t)$ *is a function of exogenous variables* $\epsilon_U, V_T, V_Y, \epsilon_M$, *thus:*

$$\Big(\epsilon_U, V_T, V_Y, \epsilon_M\Big) \perp\!\!\!\perp Z \;\Rightarrow\; M(t) \perp\!\!\!\perp Z.$$

Exogenous variables $(Z, V_T, \epsilon_T, \epsilon_U, V_Y, \epsilon_Y)$ *are multually statistically independent, and thereby* $(Z, V_T, \epsilon_T) \perp\!\!\!\perp (\epsilon_U, V_Y, \epsilon_Y)$ *holds. But according to* (55), T *is a function of exogenous variables* (Z, V_T, ϵ_T), *thus* $(T, Z) \perp\!\!\!\perp (\epsilon_U, V_Y, \epsilon_Y)$, *which also implies that* $Z \perp\!\!\!\perp (\epsilon_U, V_Y, \epsilon_Y)|T$. *Moreover, according to* (60), $Y(m)$ *is a function of exogenous variables* $(\epsilon_U, V_Y, \epsilon_Y)$, *thus:*

$$Z \perp\!\!\!\perp (\epsilon_U, V_Y, \epsilon_Y)|(T = t) \;\Rightarrow\; Z \perp\!\!\!\perp f_Y(t, m, f_U(t, \epsilon_U), V_Y, \epsilon_Y)|(T = t) \;\Rightarrow\; Z \perp\!\!\!\perp Y(m)|(T = t).$$

Appendix C Proof of Corollary C-1

Proof P-2 *We need to prove that* $P(Y(m) \leq y|T = t) = P(Y(m, t) \leq y)$ *for* $y \in supp(Y)$. *It is useful to express counterfactual* $Y(m, t)$ *as a function of exogenous variables:*

$$
\begin{aligned}
Y(m, t) &= f_Y(t, m, U(t), V_Y, \epsilon_Y)\\
&= f_Y(t, m, f_U(t, \epsilon_U), V_Y, \epsilon_Y) && (54)
\end{aligned}
$$

Moreover, exogenous variables $\epsilon_U, V_Y, \epsilon_Y, Z, V_T, \epsilon_T$ are mutually statistically independent, and, in particular:

$$\Big(\epsilon_U, V_Y, \epsilon_Y\Big) \perp\!\!\!\perp \Big(Z, V_T, \epsilon_T\Big). \tag{55}$$

According to Equations (60), we have that:

$$\begin{aligned} P(Y(m) \leq y|T=t) &\equiv P(f_Y(t,m,U,V_Y,\epsilon_Y) \leq y|T=t), \\ &\equiv P(f_Y(t,m,f_U(t,\epsilon_U),V_Y,\epsilon_Y) \leq y|f_T(Z,V_T,\epsilon_T)=t), \\ &= P(f_Y(t,m,f_U(t,\epsilon_U),V_Y,\epsilon_Y) \leq y), \\ &\equiv P(Y(t,m) \leq y), \end{aligned}$$

where the third equality comes from the independence relation (62).

Online Appendix

to

"Instrumental Variables and Causal Mechanisms: Unpacking the Effect of Trade on Workers and Voters"

Online Appendix A Data Sources

Online Appendix A.1 Election Data

We focus on federal elections (*Bundestagswahlen*) because the timing of state elections (*Landtagswahlen*) and local elections (*Kommunalwahlen*) varies widely across German regions. Federal elections took place in 1987, in December 1990 after the reunification on October 3, and in 1994, 1998, 2002, 2005, and 2009. We define the first-period outcomes as changes in the vote-share from 1987 to 1998, and second-period outcomes as changes from 1998 to 2009. Election outcomes are observed at the level of 412 districts ('*Landkreis*') in Period 2 and 322 West German districts in Period 1.

The average vote share of extreme-right parties is persistently below 5 percent in both periods. This presented a major challenge for our data collection, since official election statistics do not report all votes shares below the 5 percent minimum threshold separately by party. To extract information on extreme-right parties form this residual category, we had to contact the statistical offices of the German states and digitize some results from hard copies. By doing so, we have generated a unique data set that provides detailed insight into Germany's political constellation and allows us to create a precise measure of spatial variation in preferences also for friinge parties. This measure eventually allows us to extend existing studies on spatial variation of extreme-right activities and partisanship that were typically bound to the state level (Falk, Kuhn, and Zweimüller (2011), Lubbers and Scheepers (2001)) or limited in their time horizon (Krueger and Pischke (1997)) to a new level of detail.

Online Appendix A.2 Trade Data

Our trade data stem from the U.N. Commodity Trade Statistics Database (Comtrade). The database provides information on trade flows between country pairs, detailed by commodity type. As in Dauth et al. (2014), we express all trade flows in thousands and convert them to 2005 Euros. To merge four-digit SITC2 product codes with our three-digit industry codes, we use a crosswalk provided by Dauth et al. (2014), who themselves employ a crosswalk provided by the U.N. Statistics Division to link product categories to NACE industries. In 92 percent of the cases, commodities map unambiguously into industries. For ambiguous cases, we use national employment shares from 1978 to partition them to industries. In this way, we end up with 157 manufacturing industries (excluding fuel products), classified according to the WZ73 industry classification.

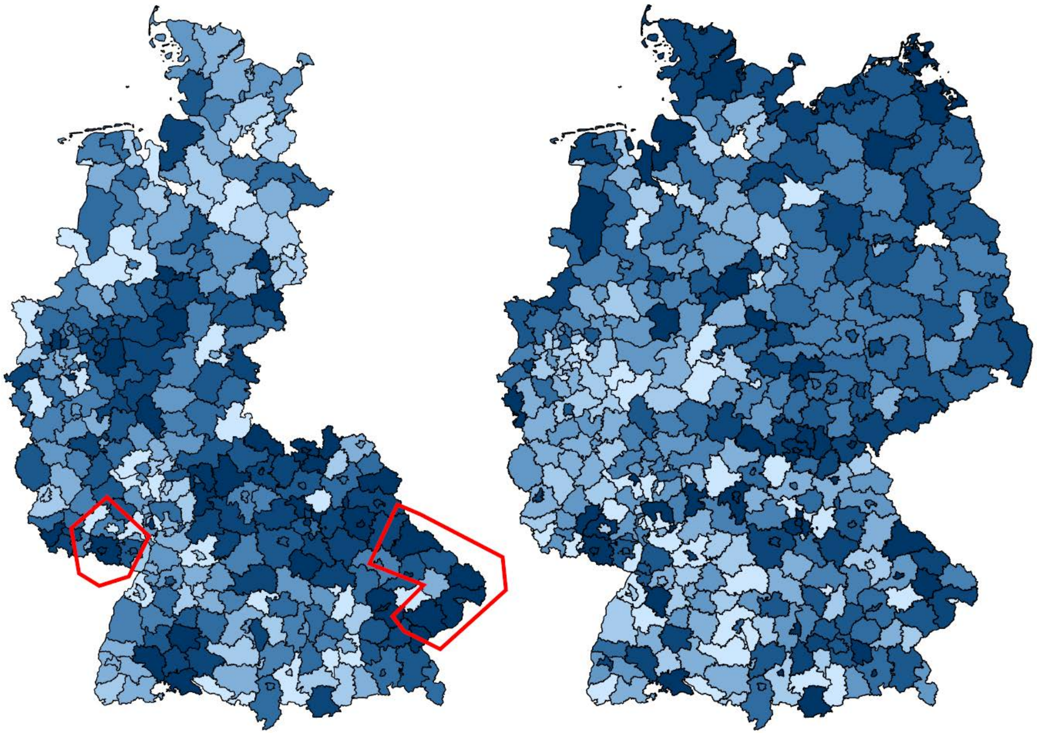
Online Appendix A.3 Labor Market Data

We obtain information on local labor markets from two different sources. Information on employment, education, and the share of foreigners stems from the Social Security records in Germany.[49] Based on the Social Security records, we calculate the trade exposure measures for local labor markets, the share of high-skilled workers (with a tertiary degree), foreign workers, workers in the automobile or chemical industry, and wages. For the years before 1999, social security data are recorded at the place of work only. After 1999, place-of-work and place-of-residence information is available.

The remaining variables are provided by the German Federal Statistical Office. These variables include the overall population, the female population share, the population share of individuals

[49] See Bender et al. (2000) for a detailed description of the data from the Institute for Employment Research (IAB). For an additional description of the regional distribution of wages across German municipalities, see Falck, Heblich, and Otto (2013)

Online Appendix Figure 1: T_{it} in 1987–1998 (Left), and 1998–2009 (Right)



Notes: Trade Shocks mapped into 322 West German counties for 1987–1998 (left) and into 408 German counties for 1998–2009 (right). The two circles enclose the regions in Palatine (on the left) and Bavaria (on the right).

of working age (aged 18 to 65), the population share of individuals older than 65, and the unemployment rate, which is calculated by dividing the number of unemployed individuals by the working-age population.

Figure A1 shows the spatial dispersion of our key regressor, T_{it}. A first observation is that there appears to be little auto-correlation in the trade exposure measure between the two periods (i.e. regions that are equally dark or light in both periods). This partly reflects the changing source of trade competition over time. While we consider trade flows from Eastern Europe and China in both periods, Eastern Europe imposes the dominant shock on German local labor markets in 1987–1998, while the shock from China dominates in the period 1998–2009 (Dauth et al., 2014). A second observation is that shocks are spatially dispersed and not clustered by state, reflecting Germany's diverse pattern of industrial production. Third, the patterns we observe are consistent with our knowledge of the spatial dimension of structural change in Germany over the past two decades. The narrative of the two circled regions in Figure A1 illustrates the the nexus of import competition, structural decline in manufacturing, and changes in voting behavior.

The circled region in the south-west of our map is Southwest-Palatine (*Südwestpfalz*), a region that was characterized by shoe and leather manufacturing firms. Increasing trade integration was a big shock to this region, centered as it was on traditional labor-intensive manufacturing industries. Today, the region–with its two main cities, Pirmasens and Zweibrücken–is considered to be one of the structurally weakest regions in West Germany; it experienced significant outmigration of young and skilled workers. Over the 1990-2006 period, Pirmasens saw a 14 percent decline in population and its unemployment rate in 2005 was at about 20 percent. A study commissioned by the Friedrich Ebert Foundation (Hafeneger and Schönfelder (2007)) investigated (among others) the case of Pirmasens and conducted interviews with local politicians to help define strategies against right-extremist parties in local parliaments. The interviews suggest that the *Republikaner*, who were represented in the city parliament, tried to mobilize voters by explicitly linking the social hardships observed to excessive globalization. In our data, Southwest-Palatine is in the top decile of negatively shocked districts in both periods. In 1987–1998, $\widehat{T}_{it} = 3.62$, while in 1998–2009, $\widehat{T}_{it} = 4.25$ in thousands of constant 2005 Euros per worker. Consistent with this, extreme-right parties increased their vote-share from 1.3 percent in 1987 to 3.45 percent in 2009.

The circled regions in in the south east of the map are located in South-Eastern Bavaria. They are bordering Austria or the Czech Republic in the so-called Dreiländereck. From the southwest to northeast, the districts are: Rottal-Inn, Passau (with the city of Passau visible in the middle), Freyung-Grafenau, Regen, and Cham. The region is known as traditional manufacturing region specialized in glass products and wood products. These labor-intensive industries were all hit hard by rising international competition which triggered a period of structural change. Today, only a few important players like Nachtmann Crystal A.G. and Schott A.G. have survived this tumult while the vast majority of small firms has disappeared. The years of structural change saw increasing unemployment and an exodus of young and skilled workers, which left the local labor market in tatters. At the same time, the region was known for right-extremist activities that attracted international attention with the near-fatal attack on Passau's police chief in 2008, which was supposedly carried out by neo-Nazis. As reported in the *New York Times* (2009), the police chief "has been known for his hard line against the extreme right, but earned the particular enmity of neo-Nazi groups after ordering the opening of the grave of a prominent former Nazi, Friedhelm Busse, after his death last July. Mr. Busse was buried with a flag bearing a swastika, which is outlawed in Germany, and the police removed the flag as evidence."

Online Appendix B Background on German Politics 1987 to 2009

Online Appendix B.1 The German Election System

Since the end of WWII, Germany has had a multiparty party system, with the two largest parties–the *Christian Democratic Union* (CDU) and the *Social Democratic Party of Germany* (SPD)–forming coalitions with either the *Free Democratic Party* (FDP) or the Greens (*Bündnis 90/Die Grünen*) during our observation period (1987 to 2009).[50] German elections are based on the principle of proportionality. The main vote, called the "second vote" (*Zweitstimme*), is being cast for parties but not for individual candidates.[51] We will exclusively focus on this party vote. The overall number of parliamentary seats is determined in proportion to a party's share of the second vote. Parties further have to surpass a 5 percent minimum threshold to be represented in federal parliament. However, this does not mean that small parties do not capture any votes. Small parties that failed to pass the 5 percent threshold still captured about 11 percent of the total votes in our election data.

[50] In this paper, we will always report votes for the CDU and its Bavarian subsection *Christian Social Union* (CSU) as combined CDU votes and refer to it as the CDU.

[51] Voters can additionally elect individual candidates on a first-past-the-post basis. Ironically, this second ballot is called the "primary vote" (*Erststimme*). In every election district, the candidate who wins the majority of primary votes is directly elected to parliament. However, electoral law ensures that this has no significant effect on the overall distribution of seats, which is determined by the second vote.

Online Appendix B.2 The Political Party Spectrum in Germany

We always classify the CDU, the SPD, the FDP, and the Greens as established parties. The conservative CDU and the social-democratic SPD are the dominant parties in Germany, in terms of both membership and votes obtained. For our period of analysis, one of those two parties was always in power. The liberal FDP participated in governments led by the CDU. The Greens are, for ideological reasons, usually the SPD's preferred coalition partner. On the extreme right of the political spectrum, three parties have regularly run in federal elections. The National Democratic Party of Germany (NPD - *Nationaldemokratische Partei Deutschlands*), founded in 1964, the Republicans (REP - *Die Republikaner*), founded in 1983, and the German People's Union (DVU - *Deutsche Volksunion*), founded in 1987 (and merged with the NPD in 2011).[52] They all follow neo-Nazi ideologies, are anti-democratic, polemicize against globalization, and agitate against immigrants and foreigners. All three have been monitored by the German Federal Office for the Protection of the Constitution (*Verfassungsschutz*). None of these extreme-right parties has ever passed the 5 percent hurdle required to enter Germany's national parliament, and it is unthinkable that any mainstream party would ever form a coalition with them (see Art (2007)). On the far left of the political spectrum, there are around 10 parties and factions that are often related with each other. Besides the left party (*Die Linke*) and its predecessors, the *Party of Democratic Socialism* (PDS) and *Labour and Social Justice The Electoral Alternative* (WASG), three branches have been dominant: Successors to the Communist Party of Germany, which had been outlawed in 1956, e.g., the *German Communist Party* (DKP) and the *Communist Party of German*y (KPD); Leninist, Stalinist, and Maoist organizations like the Marxist-Leninist Party of Germany (MLPD); and Trotskyist organizations such as the Party for Social Justice (PSG). Like the parties on the extreme right, these far-left parties are regularly monitored by either the Federal Office for the Protection of the Constitution or its state-level equivalents. We classify other parties that ran for elections but do not fit the above categories as small parties (see Falck et al. 2014).

Online Appendix B.3 Stance on Trade and Globalization

Both the large parties CDU and SPD have market-liberal as well as protectionist factions. In comparison, the CDU tends to be more market-friendly. Still, it was a government led by the SPD that implemented substantial labor market reforms in 2003-2005, amongst others decreasing employment protection, unemployment benefits, and establishing a low wage sector in Germany. The smaller FDP explicitly follows a market-liberal agenda, while the Green party focusses on environmental issues. More generally, the political left has traditionally been seen as opposing globalization and capturing the anti-globalization vote.[53] However, this is no unambiguous relationship, as the *The Economist* (2016) observes when headlining "Farewell, left versus right. The contest that matters now is open against closed." Throughout Europe, the political left has found it difficult to take a coherent position against globalization in the last two decades, often hampered by internal intellectual conflicts (Sommer 2008, Arzheimer 2009). In contrast, the right and far right successfully attended an anti-globalization agenda (Mughan et al., 2003). For the case of Germany, Sommer (2008, p. 312) argues that "in opposing globalization, the left-wing usually criticizes an unjust and profit-oriented economic world order. [It] does not reject globalization per se but rather espouses a different sort of globalization. In contrast, the solutions proposed by the

[52] In Online Appendix B.4, we provide a history of these three parties. See also comprehensive work by Stöss (2010) or Mudde (2000).

[53] To some extent this may still be the case. Che et al. (2016) for example argue that trade liberalization with China has turned American voters towards the Democrats, though it seems as if this might have not been true for the 2016 presidential elections.

extreme right keep strictly to a national framework. The extreme right's claim, therefore, that it is the only political force that opposes globalization fundamentally [...] rings true." The following excerpt from the extreme-right NPD's 'candidate manual' illustrates how Germany's far right rolls protectionist anti-globalization themes into its broader nationalistic, anti-Semitic agenda: "Globalization is a planetary spread of the capitalist economic system under the leadership of the Great Money. This has, despite by its very nature being Jewish-nomadic and homeless, its politically and military protected location mainly on the East Coast of the United States" (Grumke, 2012, p. 328).[54]

Online Appendix B.4 The Extreme-Right in West Germany

There is a strong sense of historical cultural roots and their time-persistence when it comes to explaining votes for far-right parties in Germany today. Mocan and Raschke (2014) use state-level survey aggregates from the ALLBUS, a general population survey for Germany, to show that people who live in states that had provided above-median support of the Nazi party in the 1928 elections have stronger anti-semitic feelings today. Voigtländer and Voth (2015) use the same data to show that the effects of historical antisemitic attitudes on today's political attitudes was amplified for the cohorts that grew up during Nazi Germany's indoctrination programs in 1933–1945.

Having said that, there is substantial time-variation in the popularity of the far-right in Germany. The NPD, the oldest of the three major right-wing parties we consider, was founded in 1964 as the successor to the German Reich Party (DRP). Its goal was to unite a number of fragmented far-right parties under one umbrella. Between 1966 and 1968, the NPD was elected into seven state parliaments, and in the 1969 federal election it missed the 5 percent minimum threshold by just 0.7 percentage points. Afterwards, support for the NPD declined and it took the NPD more than 25 years to re-enter state parliaments in Saxony (2004) and Mecklenburg-Western Pomerania (2006). In both states, the party got reelected in the subsequent elections, in 2009 and 2011, respectively. In 2001, the federal parliament brought in a claim to the German Constitutional Court to forbid the NPD due to its anti-constitutional program. The claim was turned down in 2003 because the NPD's leadership was infiltrated by domestic intelligence services agents, which caused legal problems. A second claim to forbid the party, filed on December 7th 2015, was denied by the constitutional judges on January 17th 2017.

The DVU was founded by publisher Gerhard Frey as an informal association in 1971. Frey published far-right newspapers such as the German National Newspaper (DNZ) and a number of books with the goal of mitigating Germany's role in WWII. His reputation as a publisher of far-right material helped Frey to become an influential player in the German postwar extreme right scene (Mudde (2000)). In 1986, Frey took it one step further starting his own far-right party German List (*Deutsche Liste*). After some name changes, the party became known as German People's Union (DVU) from 1987 on. Since its foundation, the DVU got parliamentary seats in the state assemblies of Brandenburg (1999, 2004), Bremen (1991, 1999, 2003, 2007), Schleswig-Holstein (1992), and Saxony-Anhalt (1998). In 2010, the DVU merged with the NPD.

The Republicans (Die Republikaner) were founded in 1983 as an ultraconservative breakaway from the Christian Democratic Union (CDU) and the Christian Social Union of Bavaria (CSU). Under their leader, Franz Schönhuber (who also ran as a candidate for the DVU and NPD in his later political career), the party moved further to the extreme right by propagating a xenophobic

[54] The bundling of protectionist anti-globalization themes with xenophobic content has also been noted in the 2016 U.S. presidential election, see for example *The Guardian* (2016).

view on immigrants, and particularly asylum seekers. Compared to the NPD and DVU, the Republicans were considered to be less openly extreme right which helped it secure votes from the ultraconservative clientele. The REP got parliamentary seats in Berlin's senate (1989) and the state parliament of Baden-Wuerttemberg (1992, 1996).

Online Appendix B.5 The Extreme-Right in East Germany after the Reunification

In the first decade after reunification, only the two mainstream parties, CDU and SPD, were able to establish themselves regionwide in East Germany next to the Party of Democratic Socialism (PDS), the successor of the Socialist Unity Party (SED), which had been ruling the German Democratic Republic till its collapse.

During this time smaller parties were struggling to put a party infratructure into place in East Germany. Accordingly, while all three extreme-right parties tried to establish themselves in East Germany after reunification, they did not gain major political attention until the late 1990s (Hagan, Merkens, and Boehnke, 1995). At the same time, we saw some of the worst excesses of far-right crime in East Germany in the early 1990s, when migrants' and asylum seekers' residences were set on fire and a mob of people from the neighborhood applauded. Research by Krueger and Pischke (1997) suggests that neither unemployment nor wages can explain these incidences of extreme-right-driven crime from 1991 to 1993. It is more likely that the sudden increase in the number of immigrants and asylum seekers caused these xenophobic excesses in the early 1990s.

In the mid-1990s, the initial euphoria of reunification passed and East German labor markets experienced stronger exposure to international competition. East Germany now faced almost twice as much unemployment as West Germany, and this economic malaise caused feelings of deprivation that often transformed into violent crime against immigrants. Militant right-wing groups declared "nationally liberated zones" in East Germany where foreigners were undesired. In line with that, Lubbers and Scheepers (2001) find that unemployed people have been more likely to support extreme right parties in Germany, and Falk et al. (2011) find a significant relationship between extreme-right crimes and regional unemployment levels over the years 1996–1999.[55] The story goes that the political heritage of the GDR may have preserved ethnic chauvinism, which, in in combination with subsequent economic hardship, provided a fertile ground for extreme-right parties.

[55] Note that Falk et al.'s (2011) findings do not necessarily contradict Krueger and Pischke (1997) who find no relationship between unemployment and extreme-right-driven crimes. It may very well be that the motivation for crimes changed over the 1990s.

Online Appendix C Robustness and Further Results

Online Appendix C.1 Additional Results on the Core Table 3

Online Appendix C table 2 presents the OLS results corresponding to the paper's table 3. Online Appendix C table 1 reports the coefficients on all controls in our core table 3. The initial share of manufacturing is significantly associated with increases in the extreme-right vote-share over time. In line with that, unreported specifications show that omitting the initial manufacturing share considerably increases the estimated effect of $\mathbf{T}_{it}$ on extreme-right voting. While not our focus, this relationship suggests that general structural decline and economic depression provide fertile grounds for extreme-right parties (Arzheimer, 2009). Regions with more educated workers and higher female labor force participation are less prone to shift right. Older demographics appear more prone to vote right, a finding that corroborates qualitative evidence (Stöss, 2010). Finally, high initial vote shares for extreme-right parties imply a reversion in the data, perhaps indicating cyclicality, where past swing voters to the right tend to swing back toward the mainstream.

Online Appendix C.2 Place of Work and Place of Residence Results

Here, we replicate our main result for 1999-2009 to gauge the attenuation caused by combining *place-of-residence* voting data with *place-of-work* employment data. We compare the effect of $\mathbf{T}_{rt}$, measured at the Landkreis of residence ('r'), with that of $\mathbf{T}_{it}$, measured as before at the Landkreis of work ('i').[56] Panels A and B of table 3 compare the place-of-residence with the place-of-work results. The place-of-work results for only 1999-2009 in Online Appendix C table 3 panel B are practically identical to those in column 5 of the paper's main table 3. The similarity holds if we exclude East German regions which were not contained in the first period of the sample employed in table 3. When looking at the place-of-residence in panel A, we find the same pattern in that only the vote share of extreme-right parties responds significantly to increasing trade exposure. As expected, the effect is now larger in magnitude. While we cannot extend this exercise to earlier years, table 3 does suggest that the trade effect on the extreme-right may be around 50 percent ($0.124/0.08 = 1.55$) larger when the trade shock is measured at the place of residence.

Online Appendix C.3 Labor Market Outcomes

Like Table 2 did for the voting regressions, Online Appendix C table 4 presents the OLS results corresponding to the labor market regressions in Online Appendix C table 6. Online Appendix C table 5 reports coefficients on all control variables for table 6 in the paper. .

[56] The estimation for $\mathbf{T}_{it}$ differs from that in our main exercise in two minor respects: One, in the core exercise, period 2 starts in the election year 1998, while here it starts in 1999. Two, we cannot lag employment in our instrument due to the absence of employment data at place-of-residence before 1999.

Online Appendix Table 1: Coefficients on Controls in Table 3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Turnout	CDU/CSU	SPD	FDP	Green Party	Right	Left	Small
T_{it}	0.002	-0.066	-0.009	0.119	-0.018	0.089**	-0.092	-0.024
	(1.223)	(-0.501)	(-0.073)	(1.583)	(-0.413)	(2.055)	(-0.859)	(-0.564)
Controls Specification 1:								
Empl-share manufacturing $_{-1}$	-0.000	0.023	0.002	0.009	-0.024***	0.017**	-0.010	-0.017**
	(-1.303)	(1.092)	(0.122)	(0.793)	(-2.932)	(2.407)	(-0.776)	(-2.430)
Controls Specification 2:								
Pop-share college-educated $_{-1}$	0.004***	-0.041	0.131**	-0.055	0.156***	-0.093***	-0.146**	0.049
	(2.920)	(-0.811)	(2.538)	(-1.341)	(3.530)	(-5.032)	(-2.197)	(1.544)
Pop-share foreign-born $_{-1}$	0.001	-0.205***	-0.154*	0.156***	-0.008	0.094***	0.095	0.021
	(0.358)	(-3.020)	(-1.820)	(3.820)	(-0.185)	(3.708)	(1.228)	(0.672)
Pop-share female $_{-1}$	0.011***	0.353**	-0.012	0.056	0.160	-0.262***	-0.325***	0.029
	(3.104)	(2.146)	(-0.064)	(0.534)	(1.475)	(-3.083)	(-2.602)	(0.408)
Employm-share in automotive $_{-1}$	-0.000	0.019	-0.038**	-0.001	0.030*	-0.004	-0.002	-0.004
	(-0.045)	(0.629)	(-2.047)	(-0.091)	(1.827)	(-0.353)	(-0.157)	(-0.475)
Employm-share in chemistry $_{-1}$	-0.000	0.036	-0.050***	-0.013	0.017	0.014	-0.004	-0.002
	(-0.955)	(1.214)	(-3.196)	(-0.889)	(0.915)	(0.821)	(-0.199)	(-0.189)
Employment in largest industry $_{-1}$	0.024	-1.807	2.668**	-1.159	-1.649*	0.077	0.739	1.132**
	(0.810)	(-1.090)	(1.982)	(-1.212)	(-1.704)	(0.084)	(0.569)	(2.071)
Controls Specification 3:								
Unemployment-share $_{-1}$	-0.003**	0.061	-0.034	-0.145***	-0.112***	-0.051	0.347***	-0.066***
	(-2.539)	(0.819)	(-0.431)	(-2.897)	(-2.966)	(-1.467)	(3.576)	(-2.652)
Pop-share above age 65 $_{-1}$	-0.005***	-0.113*	-0.077	-0.013	-0.002	0.079***	0.142***	-0.017
	(-3.461)	(-1.661)	(-1.137)	(-0.314)	(-0.053)	(2.598)	(3.215)	(-0.563)
Voter Turnout $_{-1}$	-0.000	0.073***	0.115***	-0.036*	-0.023	-0.016	-0.061*	-0.052***
	(-0.535)	(2.939)	(4.710)	(-1.740)	(-1.562)	(-1.638)	(-1.934)	(-3.519)
CDU/CSU Voteshare $_{-1}$	-0.025***	-0.255	-0.111	0.222	0.004	-0.635***	0.612***	0.163
	(-4.059)	(-0.987)	(-0.506)	(1.217)	(0.033)	(-4.891)	(3.227)	(1.177)
SPD Voteshare $_{-1}$	-0.010**	-0.119	-0.366*	-0.079	0.142	-0.084	0.064	0.441***
	(-2.327)	(-0.502)	(-1.946)	(-0.521)	(1.310)	(-0.993)	(0.284)	(3.561)
FDP Voteshare $_{-1}$	-0.010**	-0.293	-0.392**	-0.024	0.022	-0.089	0.373**	0.403***
	(-2.411)	(-1.334)	(-2.143)	(-0.161)	(0.218)	(-1.041)	(1.977)	(3.328)
Green Party Voteshare $_{-1}$	-0.010**	-0.081	-0.628***	-0.089	0.026	-0.076	0.440**	0.409***
	(-2.381)	(-0.373)	(-3.599)	(-0.588)	(0.262)	(-0.903)	(2.426)	(3.387)
Far-Right Voteshare $_{-1}$	-0.012***	0.120	-0.488***	-0.225	0.007	-0.098	0.359*	0.324***
	(-2.897)	(0.528)	(-2.643)	(-1.491)	(0.072)	(-1.165)	(1.667)	(2.702)
Far-Left Voteshare $_{-1}$	-0.014***	-0.349	-0.321	-0.127	0.059	-0.091	0.468**	0.359***
	(-3.060)	(-1.572)	(-1.625)	(-0.791)	(0.468)	(-1.021)	(2.338)	(2.885)
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730	730	730

Notes: T-statistics reported, standard errors are clustered at the level of 96 commuting zones, *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix Table 2: OLS Version of Table 3

	(1) Baseline OLS	(2) + Structure OLS	(3) + Industry OLS	(4) + Voting OLS	(5) +Socio OLS	(6) Standard. OLS
Δ Turnout	0.004*** (2.932)	0.003*** (2.669)	0.004*** (3.059)	0.003** (2.337)	0.003** (2.430)	0.040** (2.430)
Established Parties:						
Δ Vote Share CDU/CSU	-0.081 (-1.015)	-0.093 (-1.204)	-0.113 (-1.423)	-0.062 (-0.963)	-0.067 (-1.020)	-0.016 (-1.020)
Δ Vote Share SPD	-0.037 (-0.416)	-0.035 (-0.399)	-0.044 (-0.471)	0.061 (0.884)	0.062 (0.929)	0.005 (0.929)
Δ Vote Share FDP	0.094** (1.971)	0.114*** (2.672)	0.105** (2.398)	0.081* (1.805)	0.088** (2.097)	0.016** (2.097)
Δ Vote Share Green Party	0.046 (1.221)	0.034 (1.016)	0.063* (1.755)	0.062* (1.835)	0.068** (2.042)	0.024** (2.042)
Non-established Parties						
Δ Vote Share Extreme-Right Parties	0.038* (1.703)	0.042** (1.963)	0.036 (1.522)	-0.009 (-0.483)	-0.004 (-0.240)	-0.002 (-0.240)
Δ Vote Share Far-Left Parties	-0.108* (-1.669)	-0.105 (-1.565)	-0.109 (-1.597)	-0.138** (-2.159)	-0.153** (-2.491)	-0.039** (-2.491)
Δ Vote Share Other Small Parties	0.048 (1.586)	0.042 (1.439)	0.062** (2.186)	0.003 (0.138)	0.007 (0.259)	0.005 (0.259)
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: T-statistics reported, standard errors are clustered at the level of 96 commuting zones, *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix Table 3: Place of Work and Place of Residence

3.A: 1999–2009 data for "Place of Residence"

	(1) Turnout IV	(2) CDU/CSU IV	(3) SPD IV	(4) FDP IV	(5) Green Party IV	(6) Right IV	(7) Left IV	(8) Small IV
Period 2								
$\mathbf{T}_{rt}$	0.004	-0.123	-0.165	-0.025	0.117	0.124*	-0.029	0.102
	(1.082)	(-0.722)	(-0.815)	(-0.193)	(1.338)	(1.912)	(-0.140)	(1.251)
Period 2, West only								
$\mathbf{T}_{rt}$	0.004	-0.061	-0.069	-0.007	0.083	0.148**	-0.180	0.085
	(0.920)	(-0.342)	(-0.358)	(-0.049)	(0.822)	(2.147)	(-0.992)	(0.996)

3.B: 1999–2009 data for "Place of Work"

	(1) Turnout IV	(2) CDU/CSU IV	(3) SPD IV	(4) FDP IV	(5) Green Party IV	(6) Right IV	(7) Left IV	(8) Small IV
Period 2								
$\mathbf{T}_{it}$	0.002	-0.049	-0.182	0.093	0.109*	0.080*	-0.088	0.038
	(0.685)	(-0.424)	(-1.358)	(1.232)	(1.674)	(1.948)	(-0.569)	(0.685)
Period 2, West only								
$\mathbf{T}_{it}$	0.004	-0.059	-0.146	0.076	0.100	0.088**	-0.117	0.058
	(0.991)	(-0.496)	(-1.142)	(1.010)	(1.361)	(1.970)	(-0.864)	(0.970)

Notes: This table compares the effect of trade exposure, depending on whether **T** is measured at the *place of residence*, i.e. $\mathbf{T}_{rt}$ where 'r' replaces 'i', or as before at the place of work, $\mathbf{T}_{it}$. *Place of residence* data becomes available in 1999, i.e. one year later than the 1998 election. All specifications include identical controls to our preferred specification in Table 3, Column 5. In each panel, we also show a specification with only the 322 districts in West Germany. Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix Table 4: OLS Version of Table 6

	(1) Baseline OLS	(2) + Structure OLS	(3) + Industry OLS	(4) + Voting OLS	(5) +Socio OLS	(6) Standard. OLS
Δ Share Manufacturing Employment	-0.502***	-0.530***	-0.524***	-0.496***	-0.502***	-0.165***
	(-3.348)	(-3.613)	(-3.486)	(-3.289)	(-3.362)	(-3.362)
Δ log(Mean Manufacturing Wage)	-0.003**	-0.003**	-0.004**	-0.003**	-0.003**	-0.051**
	(-2.122)	(-2.213)	(-2.262)	(-2.094)	(-2.152)	(-2.152)
Δ log(Mean Non-Manufacturing Wage)	-0.001	-0.001	-0.001	-0.000	-0.000	-0.004
	(-0.934)	(-1.244)	(-0.853)	(-0.351)	(-0.433)	(-0.433)
Δ log(Total Employment)	-0.013***	-0.012***	-0.011**	-0.009**	-0.009*	-0.075*
	(-3.138)	(-3.066)	(-2.514)	(-2.070)	(-1.919)	(-1.919)
Δ Share Unemployment	0.089*	0.106**	0.095	0.102*	0.125***	0.068***
	(1.659)	(2.078)	(1.617)	(1.732)	(2.674)	(2.674)
Δ log(Total Population)	-0.003*	-0.002*	-0.001	-0.000	0.000	0.006
	(-1.783)	(-1.688)	(-0.807)	(-0.022)	(0.311)	(0.311)
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: T-statistics reported, standard errors are clustered at the level of 96 commuting zones, *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix Table 5: Coefficients on Controls in Table 6

	(1)	(2)	(3)	(4)	(5)	(6)
	Manuf	log(Mean Manuf. Wages)	log(Mean Non-Manuf. Wages)	log(Total Empl.)	Share Unempl.	log(Pop.)
T$_{it}$	-0.755***	-0.006***	-0.001	-0.024***	0.110*	-0.004*
	(-3.745)	(-2.592)	(-0.808)	(-3.295)	(1.694)	(-1.852)
Controls Specification 1:						
Empl-share manufacturing $_{-1}$	-0.107***	0.001***	-0.001***	-0.001	0.021***	-0.001**
	(-4.519)	(2.920)	(-5.459)	(-1.510)	(2.920)	(-2.325)
Controls Specification 2:						
Pop-share college-educated $_{-1}$	0.067	0.002	0.006***	0.010***	-0.055**	0.006***
	(0.850)	(1.631)	(8.164)	(3.876)	(-2.046)	(2.817)
Pop-share foreign-born $_{-1}$	-0.476***	-0.000	0.001	-0.013***	0.156***	-0.004**
	(-5.573)	(-0.036)	(1.206)	(-4.329)	(5.373)	(-2.202)
Pop-share female $_{-1}$	-0.062	-0.007**	0.003*	0.005	0.059	0.004
	(-0.348)	(-2.388)	(1.675)	(0.660)	(0.769)	(1.108)
Employm-share in automotive $_{-1}$	-0.024	-0.001	0.000	0.001	-0.012	0.001
	(-0.606)	(-1.524)	(1.259)	(0.965)	(-0.944)	(0.646)
Employm-share in chemistry $_{-1}$	-0.145***	-0.000	0.001*	-0.001	-0.013	0.001
	(-2.713)	(-0.790)	(1.868)	(-0.613)	(-1.179)	(0.785)
Employment in largest industry $_{-1}$	-0.527	0.018	-0.002	-0.180**	0.601	-0.030
	(-0.255)	(0.499)	(-0.069)	(-2.009)	(0.965)	(-0.740)
Controls Specification 3:						
Unemployment-share $_{-1}$	0.121	0.003	0.001	-0.010***	-0.374***	-0.013***
	(1.484)	(1.575)	(1.001)	(-3.043)	(-7.499)	(-6.321)
Pop-share above age 65 $_{-1}$	-0.036	0.002	-0.002**	-0.018***	0.108***	-0.010***
	(-0.586)	(1.420)	(-2.246)	(-6.426)	(3.134)	(-6.665)
Voter Turnout $_{-1}$	0.051***	-0.000	-0.000	0.000	-0.003	-0.001*
	(2.753)	(-0.824)	(-0.569)	(0.310)	(-0.267)	(-1.918)
CDU/CSU Voteshare $_{-1}$	-0.276	-0.005	-0.002	-0.007	0.088	-0.000
	(-1.573)	(-1.559)	(-0.663)	(-0.764)	(0.976)	(-0.014)
SPD Voteshare $_{-1}$	-0.324*	-0.005*	-0.002	-0.010	0.123	-0.001
	(-1.876)	(-1.800)	(-0.727)	(-1.027)	(1.414)	(-0.251)
FDP Voteshare $_{-1}$	-0.173	-0.006**	0.001	-0.006	0.032	-0.001
	(-0.940)	(-2.100)	(0.272)	(-0.614)	(0.362)	(-0.238)
Green Party Voteshare $_{-1}$	-0.487***	-0.002	-0.003	-0.011	0.091	0.002
	(-2.711)	(-0.493)	(-1.143)	(-1.097)	(0.976)	(0.372)
Far-Right Voteshare $_{-1}$	-0.269	-0.006	-0.001	-0.013	0.231**	-0.003
	(-1.152)	(-1.591)	(-0.238)	(-1.010)	(2.015)	(-0.520)
Far-Left Voteshare $_{-1}$	-0.383**	-0.011***	-0.004	-0.020**	0.150	-0.001
	(-2.069)	(-2.888)	(-1.545)	(-2.004)	(1.552)	(-0.179)
Period-by-region FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	730	730	730	730	730	730

Notes: T-statistics reported, standard errors are clustered at the level of 96 commuting zones, *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix D Constructing Gravity Residuals

Gravity-residuals can be obtained from the residuals of the regression

$$log(EX_{djt}^{CE-O}) - log(EX_{djt}^{G-O}) = \alpha_d + \alpha_j + \epsilon_{djt}^{IM}, \tag{56}$$

where $log(EX_{djt}^{CE-O})$ are industry j's log export values from China and Eastern Europe to destination market d, $log(EX_{djt}^{G-O})$ are German industries' exports to the same countries, α_d are destination-market and α_j are industry-fixed effects.[57] ϵ_{djt}^{IM} thus captures CE's competitive advantage over Germany at time t in destination market d and industry j. Averaging residuals ϵ_{djt}^{IM} over destination markets d and taking first differences provides a measure for overall changes in CE's comparative advantage over time. Exponentiating this term and multiplying it with Germany's start-of-period imports from CE gives rise to $\Delta IM_{Gjt}^{grav} = IM_{Gjt-1} \times exp^{\bar{\epsilon}_{jt}^{IM} - \bar{\epsilon}_{jt-1}^{IM}}$, which is a counterfactual measure of changes in German industries' import exposure that is solely driven by CE's increasing comparative advantage.

Conversely, ΔEX_{Gjt} increases due to better access to the CE markets and to German-specific supply conditions. While German-specific supply conditions will affect German exports in general, the relative attractiveness of CE markets over other export destinations should be independent of German-specific effects. Thus, changes in German industries' exports to China and Eastern Europe in relation to German industries' exports to other countries O provides an exogenous measure ΔEX_{Git}^{grav} for ΔEX_{Gjt}. It can be obtained from the residuals of the regression

$$log(EX_{djt}^{G-CE}) - log(EX_{djt}^{G-O}) = \alpha_d + \alpha_j + \epsilon_{djt}^{EX}, \tag{57}$$

where $log(EX_{djt}^{G-CE})$ are industry j's log export values from Germany to China and Eastern Europe, $log(EX_{djt}^{G-O})$ are German industries' exports to other countries, and α_d and α_j are again destination-country and industry-fixed effects. ϵ_{djt}^{EX} now captures CE's relative attractiveness over other sales markets at time t in destination market d and industry j. Averaging residuals ϵ_{djt}^{EX} over destination markets d and taking first differences provides a measure for overall changes in the attractiveness of Chinese and Eastern European sales markets over time. Exponentiating this term and multiplying it with Germany's start-of-period exports to CE gives rise to $\Delta EX_{Gjt}^{grav} = EX_{Gjt-1} \times exp^{\bar{\epsilon}_{jt}^{EX} - \bar{\epsilon}_{jt-1}^{EX}}$, which is a counterfactual measure of changes in German industries' export exposure that is solely driven by CE's increasing attractiveness as sales market.

[57] Since many CE countries did not report trade data in the late 1980s and early 1990s, we use imports *from* CE and Germany reported *by* other countries O to measure Germany's and CE's exports to O.

Online Appendix E Subsample Results for the Effect of Trade T on Labor M and Voting Y

Panel A of table 6 reports the same six labor market outcomes plus their principal components as table 6, estimated separately for Period 1 (1987–1998), and Period 2 (1998–2009), as well as for West Germany only in Period 2. Panel B of table 6 similarly decomposes the same eight political outcomes as reported in table 3. The sample sizes are 322, 408, and 322 respectively.

The discussion in section 2.5 suggests that the effect of trade shocks on labor markets should be more pronounced in the second period, when companies were more flexible to react. We found some evidence for this pattern in the individual results in table 4. This motivates us to decompose the effect of trade exposure on local labor markets by period in this section. Panel A of table 6 reports the same six labor market outcomes plus their principal components as table 6, estimated separately for Period 1 (1987–1998), and Period 2 (1998–2009), as well as for West Germany only in Period 2. Panel B of table 6 similarly decomposes the same eight political outcomes as reported in table 3. The sample sizes are 322, 408, and 322 respectively.

Comparing the three sub-panels of panel A shows evidence for increasing flexibility in labor markets between Period 1 and Period 2. This is nicely reflected by the core result for manufacturing employment in column 1 and by the second principal component of labor market disturbances reported in column 8 of Panel A. In period 1, import competition also has counter-intuitive effects on wages, hinting at regulatory rigidities. The observed contrast between periods is not driven by the inclusion of East German regions in period 2. In fact, the contrast between the two periods is more pronounced once we focus on West Germany. Panel B shows that voting responses to trade were strongest when labor markets were least regulated. Combining the evidence, table 6 suggests that trade exposure had the biggest effect on both voting and labor market disturbances in the second period in West Germany, i.e. when labor markets were most deregulated and subject to market forces. We interpret this symmetry as “reduced form evidence” for the important role of labor markets as mediators in the transmission from trade shocks to voting responses. However, without additional econometric structure, it is not possible to infer on the causality of the labor market mechanisms.

Online Appendix Table 6: Decomposing the Results by Period and Imports vs Exports

6.A: Labor Market Outcomes/Mediators

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Manuf	log(mean Manuf. Wages)	log(Non-Manuf. Wages)	log(Total Empl.)	Share Unempl.	log(Pop.)	1st Principal Comp'nt	2nd Prinicip. Comp'nt
Period 1								
T_{it}	-0.324	0.005*	-0.000	-0.027	0.079	-0.006	-0.003	-0.252
	(-0.896)	(1.662)	(-0.011)	(-1.600)	(0.883)	(-1.341)	(-0.068)	(-1.435)
Period 2								
T_{it}	-0.400**	0.001	-0.002	-0.011	-0.079	-0.001	-0.017	-0.141*
	(-2.350)	(0.433)	(-1.144)	(-1.311)	(-1.112)	(-0.502)	(-0.441)	(-1.651)
Period 2, West only								
T_{it}	-0.572***	-0.001	-0.002	-0.019**	-0.055	-0.002	-0.032	-0.230**
	(-3.069)	(-0.337)	(-0.945)	(-1.990)	(-0.823)	(-0.905)	(-0.742)	(-2.453)

6.B: Final Voting Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Turnout	CDU/CSU	SPD	FDP	Greens	Right	Left	Small
Period 1								
T_{it}	0.000	-0.298	0.320	0.013	-0.003	-0.025	-0.001	-0.007
	(0.013)	(-1.159)	(1.558)	(0.150)	(-0.030)	(-0.243)	(-0.041)	(-0.105)
Period 2								
T_{it}	0.000	-0.115	-0.173	0.076	0.081	0.071*	0.058	0.003
	(0.080)	(-0.704)	(-1.072)	(0.821)	(1.142)	(1.696)	(0.360)	(0.044)
Period 2, West only								
T_{it}	0.002	-0.095	-0.161	0.083	0.110	0.084**	-0.023	0.001
	(0.514)	(-0.542)	(-0.987)	(0.886)	(1.342)	(2.078)	(-0.187)	(0.018)

Notes: The table reports subsample estimations. Panel A reports on the same eight labor market outcomes in table 6. Panel B reports on the same eight political outcomes in table 3. Every result reported in table 6 is from a TSLS estimation that breaks treatment into separate import competition and export access effects, instrumented with Z_{it}^{IM} and Z_{it}^{EX}, defined in (2). Every panel additional reports the results for three separate sub-samples: period 1 (1987–1998) and period 2 (1998–2009), and period 2 without the 86 East German districts. The sample sizes are 322, 408, and 322 respectively. All specifications include region fixed effects. Standard errors are clustered at the level of 96 commuting zones. *** p<0.01, ** p<0.05, * p<0.1.

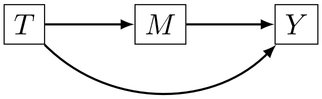
Online Appendix F Mediation Model without Confounding Variables

The simplest mediation model consists of three observed variables T, M, Y and three statistically independent error terms $\epsilon_T, \epsilon_M, \epsilon_Y$. Causal relations are defined by the following equations:

$$T = f_T(\epsilon_T)\ ,\ M = f_M(T, \epsilon_M)\ ,\ Y = f_Y(T, M, \epsilon_Y). \tag{58}$$

Input variables of functions f_T, f_M, f_Y are said to cause their respective output variables. Thus T causes M and Y while M causes Y. Neither functions f_T, f_M, f_Y nor error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ are observed. We use the notations $supp(T), supp(M), supp(Y)$ for the support of T, M, Y respectively. Figure 2 displays Model (65) as a *Directed Acyclic Graph* (DAG).

Online Appendix Figure 2: Mediation Model without Confounding Variables



Fixing is a key concept in causal analysis. Fixing is defined as the causal operation that assigns a value to an argument of a structural equation. It is used to generate counterfactual variables. For instance, Model (65) renders three counterfactual variables. The counterfactual mediator $M(t)$ is generated by fixing the argument T of function f_M to a value $t \in supp(T)$, that is, $M(t) = f_M(t, \epsilon_M)$. The counterfactual outcome $Y(t)$ for T fixed at t is given by $Y(t) = f_Y(t, M(t), \epsilon_Y)$ and the counterfactual outcome Y when T is fixed at t and M is fixed at m is given by $Y(t, m) = f_Y(t, m, \epsilon_Y)$. We refer to Heckman and Pinto (2015a) for a detailed discussion on the fixing operator, counterfactual outcomes and causal models.

We are interested in identifying the effect of T on outcome Y, but most importantly, we are interested in identifying the mechanism M though which T causes Y. This task is often referred to as mediation analysis and it requires the identification of all three counterfactual variables $Y(t)$, $M(t)$, $Y(m, t)$. Robins and Greenland (1992) examine the case of a binary treatment $supp(T) = \{t_0, t_1\}$ and define three primary causal parameters in mediation analysis: the *total*, *direct* and *indirect* effects.

$$\begin{array}{llcll}
\text{Total Eff. :} & ATE & = & E(Y(t_1) - Y(t_0)) & \equiv E(Y(t_1, M(t_1)) - Y_i(t_0, M(t_0))), \\
\text{Direct Eff. :} & ADE(t) & = & E(Y(t_1, M(t)) - Y(t_0, M(t))) & \equiv \int E\Big(Y(t_1, m) - Y(t_0, m)\Big) dF_{M(t)}(m), \\
\text{Indirect Eff. :} & AIE(t) & = & E(Y(t, M(t_1)) - Y(t, M(t_0))) & \equiv \int E\Big(Y(t, m)\Big)\Big[dF_{M(t_1)}(m) - dF_{M(t_0)}(m)\Big],
\end{array}$$

where $F_{M(t)}(m)$ stand for the cumulative probability distribution of counterfactual mediator $M(t)$. ATE is the average causal effect of T on Y. $ADE(t)$ is the causal effect of T on Y when we hold the distribution of M fixed at $M(t)$. $AIE(t)$ is the causal effect of T on Y induced by the change in the distribution of the mediator M. The total effect ATE can be expressed as the sum of direct and indirect effect as:

$$\begin{aligned}
ATE &= E(Y(t_1, M(t_1)) - Y_i(t_0, M(t_0))) \\
&= \Big(E(Y(t_1, M(t_1))) - E(Y(t_0, M(t_1)))\Big) + \Big(E(Y(t_0, M(t_1)) - Y_i(t_0, M(t_0)))\Big) = ADE(t_1) + AIE(t_0) \\
&= \Big(E(Y(t_1, M(t_1))) - E(Y(t_1, M(t_0)))\Big) + \Big(E(Y(t_1, M(t_0)) - Y_i(t_0, M(t_0)))\Big) = AIE(t_1) + ADE(t_0).
\end{aligned}$$

Model (65) has no confounding variables. That is to say that model (65) assumes no unobserved variable that jointly causes T, M and Y. This implies that variables T, M are independent of counterfactual outcomes, that is, $T \perp\!\!\!\perp \big(Y(t), M(t)\big)$ and $M \perp\!\!\!\perp Y(t,m)$. Indeed, $T = f_T(\epsilon_T)$ depends only on ϵ_T, $M(t) = f_M(t, \epsilon_M)$ and $Y(t) = f_Y(t, M(t), \epsilon_Y)$ only depend on ϵ_M, ϵ_Y. But ϵ_T is independent of ϵ_M, ϵ_Y. Thus we can write:

$$\big(\epsilon_Y, \epsilon_M\big) \perp\!\!\!\perp \epsilon_T \;\Rightarrow\; \big(f_Y(t, f_M(t, \epsilon_M), \epsilon_Y), f_M(t, \epsilon_M)\big) \perp\!\!\!\perp f_T(\epsilon_T) \;\Rightarrow\; \big(Y(t), M(t)\big) \perp\!\!\!\perp T$$

On the other hand, $Y(t,m) = f_Y(t, m, \epsilon_Y)$ only depends on ϵ_Y. Thus we can write:

$$\epsilon_Y \perp\!\!\!\perp \big(\epsilon_M, \epsilon_T\big) \;\Rightarrow\; f_Y(t, m, \epsilon_Y) \perp\!\!\!\perp f_M(f_T(\epsilon_T), \epsilon_M) \;\Rightarrow\; Y(t,m) \perp\!\!\!\perp M.$$

A substantial literature on mediation analysis assumes no confounding variables. This literature often evokes the Sequential Ignorability Assumption of Imai et al. (2010). Online Appendix G shows that Model (65) also implies Sequential Ignorability.

If the independence relations $T \perp\!\!\!\perp \big(Y(t), M(t)\big)$ and $M \perp\!\!\!\perp Y(t,m)$ hold, then we are able to express average counterfactual outcomes in terms of conditioned expectations from observed data. We illustrate this fact for the counterfactual outcome $Y(t)$. The observed outcome Y can be expressed as:

$$Y = \sum_{t \in supp(T)} Y(t) \cdot \mathbf{1}[T = t],$$

where $\mathbf{1}[T = t]$ is an indicator function that takes value one if $T = t$ and zero otherwise. If $T \perp\!\!\!\perp Y(t)$ holds then $E(Y(t)) = E(Y(t)|T = t)$ also holds and we can express $E(Y(t))$ as:

$$E(Y(t)) = E(Y(t)|T = t) = E\left(\sum_{t \in supp(T)} Y(t) \cdot \mathbf{1}[T = t] | T = t\right) = E(Y|T = t).$$

The expectation $E(Y|T = t)$ can be evaluated from observed data and thereby $E(Y(t))$ is said to be identified.

Online Appendix G The Sequential Ignorability Assumption

A large literature on mediation analysis relies on the Sequential Ignorability Assumption **A-1** of Imai et al. (2010) to identify mediation effects.

Assumption A-1 *Sequential Ignorability (Imai et al., 2010):*

$$\big(Y(t',m), M(t)\big) \perp\!\!\!\perp T|X \tag{59}$$

$$Y(t',m) \perp\!\!\!\perp M(t)|(T,X), \tag{60}$$

where X denotes pre-intervention variables that are not caused by T, M and Y such that $0 < P(T = t|X) < 1$ and $0 < P(M(t) = m|T = t, X) < 1$ holds for all $x \in supp(X)$ and $m \in supp(M)$.

Under Sequential Ignorability **A-1**, it is easy to show that the distributions of conterfactual variables are identified by $P(Y(t,m)|X) = P(Y|X, T = t, M = m)$ and $P(M(t)|X) = P(M|X, T = t)$ and thereby the mediating causal effects can be expressed as:

$$ADE(t) = \int \left(\big(E(Y|T = t_1, M = m, X = x) - E(Y|T = t_0, M = m, X = x, X = x)\big) dF_{M|T=t,X=x}(m) \right) dF_X(x) \tag{61}$$

$$AIE(t) = \int \left(E(Y|T = t, M = m, X = x) \Big[dF_{M|T=t_1,X=x}(m) - dF_{M|T=t_0,X=x}(m) \Big] \right) dF_X(x). \tag{62}$$

Imai, Tingley, Keele and Yamamoto offer a substantial line of research that explores the identifying properties of Sequential Ignorability Assumption **A-1**. See Imai et al. (2011a) for a comprehensive discussion of the benefits and limitations of the sequential ignorability assumption.

The main critics of Sequential Ignorability **A-1** is that it does not hold under the presence of either *Confounders* or *Unobserved Mediators* (Heckman and Pinto, 2015b).

The independence relation (66) assumes that T is exogenous conditioned on X. There exists no unobserved variable that causes T and Y or T and M. For instance, the Sequential Ignorability **A-1** holds for the model defined in (65) because:

$$(\epsilon_Y, \epsilon_M) \perp\!\!\!\perp \epsilon_T \Rightarrow \quad \big(f_Y(t', m, \epsilon_Y), f_M(t, \epsilon_M)\big) \perp\!\!\!\perp f_T(\epsilon_T) \Rightarrow \quad \big(Y(t', m), M(t)\big) \perp\!\!\!\perp T. \tag{63}$$

$$\epsilon_Y \perp\!\!\!\perp \epsilon_M|\epsilon_T \Rightarrow \quad f_Y(t', m, \epsilon_Y) \perp\!\!\!\perp f_M(t, \epsilon_M)|f_T(\epsilon_T) \Rightarrow \quad Y(t', m) \perp\!\!\!\perp M(t)|T, \tag{64}$$

where the initial independence relation in (70) and (71) comes from the independence of error terms.

This assumption is expected to hold in experimental data when treatment T is randomly assigned. The independence relation (67) assumes that M is exogenous conditioned on X and T. It assumes that no confounding variable causing M and Y. Sequential Ignorability **A-1** is an extension of the Ignorability Assumption of Rosenbaum and Rubin (1983) that also assumes that a treatment T is exogenous when conditioned on pre-treatment variables. Robins (2003); Petersen, Sinisi, and Van der Laan (2006); Rubin (2004) state similar identifying criteria that assume no confounding variables. Those assumptions are not testable.

Figure 7 in the paper reveals that Sequential Ignorability **A-1** assumes that: (1) the confounding variable V is observed, that is, the pre-treatment variables X; and (2) that there is no unobserved mediator U. This assumption is unappealing for many because it solves the identification problem generated by confounding variables by assuming that those do not exist (Heckman, 2008).

Consider a change in the treatment variable T denoted by $\Delta(t) = t_1 - t_0$. The Direct and indirect effects can be expressed by:

$$\begin{aligned}
ADE(t') &= \Big(\lambda_{YT} \cdot t_1 + \lambda_{YM} \cdot E(M(t'))\Big) - \Big(\lambda_{YT} \cdot t_0 + \lambda_{YM} \cdot E(M(t'))\Big) \\
\therefore ADE &= \lambda_{YT} \cdot \Delta(t) \qquad (65) \\
\text{and } AIE(t') &= \Big(\lambda_{YT} \cdot t' + \lambda_{YM} \cdot E(M(t_1))\Big) - \Big(\lambda_{YT} \cdot t' + \lambda_{YM} \cdot E(M(t_0))\Big) \\
&= \Big(\lambda_{YT} \cdot t' + \lambda_{YM}\lambda_M \cdot t_1\Big) - \Big(\lambda_{YT} \cdot t' + \lambda_{YM}\lambda_M \cdot t_0\Big) \\
\therefore AIE &= \lambda_{YM} \cdot \lambda_M \cdot \Delta(t) \qquad (66)
\end{aligned}$$

Online Appendix H Identification of Causal Parameters

The linear mediation model we investigate can be fully described by the following equations:

$$\text{Instrumental Variable } Z = \epsilon_Z, \tag{67}$$
$$\text{Treatment } T = \xi_Z \cdot Z + \xi_V \cdot V_T + \epsilon_T, \tag{68}$$
$$\text{Unobserved Mediator } U = \zeta_T \cdot T + \epsilon_U, \tag{69}$$
$$\text{Observed Mediator } M = \varphi_T \cdot T + \varphi_U \cdot U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{70}$$
$$\text{Outcome } Y = \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V_Y + \epsilon_Y, \tag{71}$$
$$\text{Exogenous Variables} \quad Z, V_T, V_M, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y \text{ are statistically independent variables,} \tag{72}$$
$$\text{Scalar Coefficients} \quad \xi_Z, \xi_V, \zeta_T, \varphi_T, \varphi_U, \delta_Y, \delta_T, \beta_T, \beta_M, \beta_U, \beta_V \tag{73}$$
$$\text{Unobserved Variables} \quad V_T, V_M, U, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y. \tag{74}$$

We assume that all variables have mean zero. This assumption does not incur in less of generality, but simplify notation as intercepts can be suppressed.

We first eliminate the unobserved mediator U from Equations (77)–(78) by iterated substitution. Equations (78)–(78) are then expressed as:

$$M = (\varphi_T + \varphi_U \zeta_T) \cdot T + \varphi_U \cdot \epsilon_U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{75}$$
$$Y = (\beta_T + \beta_U \zeta_T) \cdot T + \beta_M \cdot M + \beta_U \cdot \epsilon_U + \beta_V \cdot V_Y + \epsilon_Y. \tag{76}$$

We use the following transformation of parameters to save on notation:

$$\widetilde{\varphi}_T = \varphi_T + \varphi_U \zeta_T, \tag{77}$$
$$\widetilde{\beta}_T = \beta_T + \beta_U \zeta_T, \tag{78}$$
$$\widetilde{U} = \epsilon_U. \tag{79}$$

We use equations (82)–(86) to simplify Model (74)–(78) into the following equations:

$$\text{Instrumental Variable } Z = \epsilon_Z, \tag{80}$$
$$\text{Treatment } T = \xi_Z \cdot Z + \xi_V \cdot V_T + \epsilon_T, \tag{81}$$
$$\text{Observed Mediator } M = \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{82}$$
$$\text{Outcome } Y = \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \tag{83}$$

In this linear model, the counterfactual outcomes $M(t), Y(t), Y(m), Y(m,t)$ are given by:

$$
\begin{aligned}
M(t) &= \widetilde{\varphi}_T \cdot t + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, && (84)\\
Y(m) &= \widetilde{\beta}_T \cdot T + \beta_M \cdot m + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. && (85)\\
Y(t,m) &= \widetilde{\beta}_T \cdot t + \beta_M \cdot m + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. && (86)\\
Y(t) &= \widetilde{\beta}_T \cdot t + \beta_M \cdot M(t) + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \\
&= (\widetilde{\beta}_T + \beta_M \widetilde{\varphi}_T) \cdot t + (\beta_U + \beta_M \varphi_U) \cdot \widetilde{U} + (\beta_V + \beta_M \delta_Y) \cdot V_Y + \beta_M \delta_T \cdot V_T + \beta_M \cdot \epsilon_M + \epsilon_Y. && (87)
\end{aligned}
$$

We claim that the coefficients associated with unobserved variables $V_T, \widetilde{U}, V_Y$ may only be identified up a linear transformation. Consider the coefficients δ_T, β_V that multiply the unobserved variable V_T in Equations (88) and (89) respectively. Suppose a linear transformation that multiplies V_T by a constant $\kappa \neq 0$. The model would remain the same if coefficients δ_T, β_V were divided by the same constant κ. This is a typical fact in the literature of linear factor models. We solve this non-identification problem by impose that each unobserved variable $V_T, \widetilde{U}, V_Y$ has unit variance:

$$
\text{var}(V_T) = \text{var}(\widetilde{U}) = \text{var}(V_Y) = 1. \qquad (88)
$$

Assumption (95) is typically termed as *anchoring* of unobserved factors in the literature of factor analysis. This assumption does not incur in any loss of generality for the identification of direct, indirect or total causal effects of T (and M) on Y as expressed in the following section.

Online Appendix H.1 Defining Causal Parameters

The literature of mediation analysis term relevant causal parameters as:

- Total Effect of T on Y, that is, $\frac{dE(Y(t))}{dt}$.
- Direct Effect of T on Y, that is $\frac{\partial E(Y(t,m))}{\partial t}$.
- Effect of M on Y, that is, $\frac{dE(Y(m))}{dm}$.
- Effect of T on M, that is, $\frac{dE(M(t))}{dt}$.
- Indirect Effect of T on Y, that is $\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}$.

According to the counterfactual variables in (91)–(94), these causal effects are given by:

$$
\begin{aligned}
\text{Total Effect of } T \text{ on } Y: \frac{dE(Y(t))}{dt} &= \widetilde{\varphi}_T \cdot \beta_M + \widetilde{\beta}_T. && (89)\\
\text{Direct Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial t} &= \widetilde{\beta}_T. && (90)\\
\text{Effect of } M \text{ on } Y: \frac{dE(Y(m))}{dm} &= \beta_M. && (91)\\
\text{Effect of } T \text{ on } M: \frac{dE(M(t))}{dt} &= \widetilde{\varphi}_T. && (92)\\
\text{Indirect Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt} &= \beta_M \cdot \widetilde{\varphi}_T. && (93)
\end{aligned}
$$

Online Appendix H.2 Identifying Equations

Model (87)–(90) can be conveniently expressed in matrix notation. In Equation (101) we define $\mathbf{X} = [Z, T, M, Y]'$ as the vector of observed variables, $\mathbf{V} = [V_T, V_Y, \widetilde{U}]'$ as the vector of unobserved confounding variables, and $\boldsymbol{\epsilon} = [\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y]'$ as the vector of exogenous error terms. According to (79), the random vectors $\mathbf{V}$ and $\boldsymbol{\epsilon}$ are independent, that is, $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$. We use $\mathbf{K}$ in (101) for the matrix of parameters that multiply $\mathbf{X}$ and $\mathbf{A}$ for the matrix of parameters that multiply $\mathbf{V}$.

$$\mathbf{X} = \begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} V_T \\ V_Y \\ \widetilde{U} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ \xi_V & 0 & 0 \\ \delta_Y & \delta_Y & \varphi_U \\ 0 & \beta_V & \beta_U \end{bmatrix}. \tag{94}$$

Using the notation in (101), we can express the linear system (87)–(90) as following:

$$\underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}}_{\mathbf{K}} \cdot \underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ \xi_V & 0 & 0 \\ \delta_T & \delta_Y & \varphi_U \\ 0 & \beta_V & \beta_U \end{bmatrix}}_{\mathbf{A}} \cdot \underbrace{\begin{pmatrix} V_T \\ V_Y \\ \widetilde{U} \end{pmatrix}}_{\mathbf{V}} + \underbrace{\begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}}_{\boldsymbol{\epsilon}}, \tag{95}$$

$$\mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon}. \tag{96}$$

The coefficients in matrices $\mathbf{K}$, $\mathbf{A}$ are identified through the covariance matrices of observed variables. We use $\Sigma_{\mathbf{X}} = \text{cov}(\mathbf{X}, \mathbf{X})$ for the covariance matrix of observed variables $\mathbf{X}$, and $\Sigma_{\boldsymbol{\epsilon}} = \text{cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ for the vector of error terms $\boldsymbol{\epsilon}$. $\Sigma_{\boldsymbol{\epsilon}}$ is a diagonal matrix due to statistical independence of error terms. We also use $\Sigma_{\mathbf{V}} = \text{cov}(\mathbf{V}, \mathbf{V})$ for the covariance of unobserved variables $\mathbf{V}$. The unobserved variables in $\mathbf{V}$ are statistically independent and have unit variance (95), thus $\Sigma_{\mathbf{V}} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Moreover, $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$ implies that $\text{cov}(\mathbf{V}, \boldsymbol{\epsilon}) = \mathbf{0}$, where $\mathbf{0}$ is a matrix of elements zero.

Equation (106) determines the relation between the covariance matrices of observed and unobserved variables:

$$\begin{aligned} \mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon} &\Rightarrow (\mathbf{K} - \mathbf{I})\,\mathbf{X} = \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon}, & (97) \\ &\Rightarrow (\mathbf{K} - \mathbf{I})\,\Sigma_{\mathbf{X}}\,(\mathbf{K} - \mathbf{I})' = \mathbf{A}\Sigma_{\mathbf{V}}\mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}, & (98) \\ &\Rightarrow (\mathbf{K} - \mathbf{I})\,\Sigma_{\mathbf{X}}\,(\mathbf{K} - \mathbf{I})' = \mathbf{A}\mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}, & (99) \end{aligned}$$

where the second equation is due to $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$ and the third equations comes from $\Sigma_{\mathbf{V}} = \mathbf{I}$.

Equation (106) generates ten equalities. Four equalities are due to the diagonal of the covariance matrices $(\mathbf{K} - \mathbf{I})\,\Sigma_{\mathbf{X}}\,(\mathbf{K} - \mathbf{I})'$ and $\mathbf{A}\mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}$ in (106). The remaining six equalities from the off-diagonal relation of the covariance matrices in (106).

The diagonal elements of $\Sigma_{\boldsymbol{\epsilon}}$ are the variances of the error terms $\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y$. Thereby each diagonal equation generated by (106) adds one unobserved term to the system of quadratic equations. The point-identification of the model coefficients arises from the six off-diagonal equations generated by (106). Those are listed below:

$$\operatorname{cov}(Z,T) - \operatorname{cov}(Z,Z)\cdot \xi_Z = 0 \tag{100}$$
$$\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\cdot \widetilde{\varphi}_T = 0 \tag{101}$$
$$\operatorname{cov}(Z,Y) - \operatorname{cov}(Z,M)\cdot \beta_M - \operatorname{cov}(Z,T)\cdot \widetilde{\beta}_T = 0 \tag{102}$$
$$\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\cdot \widetilde{\beta}_T - \operatorname{cov}(T,M)\cdot \beta_M = 0 \tag{103}$$
$$\operatorname{cov}(M,Y) - \operatorname{cov}(T,M)\cdot \widetilde{\beta}_T - \operatorname{cov}(M,M)\cdot \beta_M = \beta_U\cdot\varphi_U + \beta_V\cdot\delta_Y \tag{104}$$
$$\operatorname{cov}(T,M) - \operatorname{cov}(T,T)\cdot \widetilde{\varphi}_T = \delta_T\cdot\xi_V \tag{105}$$

Simple manipulation of Equations (107)–(112) generate the identification of the following parameters:

$$\xi_Z = \frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,Z)} \qquad \text{from Eq.(107)} \tag{106}$$
$$\widetilde{\varphi}_T = \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} \qquad \text{from Eq.(108)} \tag{107}$$
$$\beta_M = \frac{\operatorname{cov}(Z,T)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)} \qquad \text{from Eqs.(109)–(110)} \tag{108}$$
$$\widetilde{\beta}_T = \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \qquad \text{from Eqs.(109)–(110)} \tag{109}$$
$$\beta_U\cdot\varphi_U + \beta_V\cdot\delta_Y = \operatorname{cov}(M,Y) - \operatorname{cov}(M,M)\cdot\beta_M - \operatorname{cov}(T,M)\cdot\widetilde{\beta}_T \qquad \text{from Eq.(111)} \tag{110}$$
$$\delta_T\cdot\xi_V = \frac{\operatorname{cov}(T,M)\operatorname{cov}(Z,M) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,M)} \qquad \text{from Eq.(112)} \tag{111}$$

Moreover, if we divide Equation (109) by $\operatorname{cov}(Z,T)$ we obtain:

$$\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}\cdot\beta_M - \frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}\cdot\widetilde{\beta}_T = 0 \tag{112}$$
$$\Rightarrow \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \widetilde{\varphi}_T\cdot\beta_M - \widetilde{\beta}_T = 0 \tag{113}$$
$$\Rightarrow \widetilde{\varphi}_T\cdot\beta_M + \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}. \tag{114}$$

The four causal of interest parameters defined in (96)–(99) are respectively identified by Equations (114), (115), (116) and (121):

$$\frac{dE(M(t))}{dt} = \widetilde{\varphi}_T = \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}, \tag{115}$$
$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\operatorname{cov}(Z,Y)\operatorname{cov}(T,T) - \operatorname{cov}(Y,T)\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,M)\operatorname{cov}(T,T) - \operatorname{cov}(M,T)\operatorname{cov}(Z,T)}, \tag{116}$$
$$\frac{\partial E(Y(t,m))}{\partial t} = \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}, \tag{117}$$
$$\frac{dE(Y(t))}{dt} = \widetilde{\varphi}_T\cdot\beta_M + \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}. \tag{118}$$

Next section explains that each causal effect (122)–(125) can be evaluated by standard Two-stage Least Squares regressions.

Online Appendix I Estimation of Causal Parameters

Our goal is to show that the four causal parameters listed in Equations (122)–(125) can be estimated using the standard Two-stage Least Square (2SLS) estimator. We revise the standard equations of the 2SLS estimators for sake of completeness.

Equations (126)–(127) present the first and stages of a generic 2SLS regression in which T stands for the endogenous variable, Z is the instrumental variable and Y is the targeted outcome.

$$\text{First Stage: } T = \kappa_1 + \beta_1 \cdot Z + \epsilon_1, \tag{119}$$

$$\text{Second Stage: } Y = \kappa_2 + \beta_2 \cdot T + \epsilon_2. \tag{120}$$

The 2SLS estimator relies on the assumptions that the instrument Z is statistically independent of the term ϵ_2 while T is not. It is well-known that the 2SLS estimator $\hat{\beta}_2$ is given by the ratio of the sample covariances $\text{cov}(Z, Y)$ and $\text{cov}(Z, T)$. Moreover $\hat{\beta}_2$ is a consistent estimator of parameter β_2 :

$$\text{plim}(\hat{\beta}_2) = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, T)} = \beta_2. \tag{121}$$

Consider the inclusion of additional covariates X in both stages of the 2SLS method. Variables X in (129)–(130) play the role of control covariates in the first stage and second stages of the 2SLS estimator. Control covariates X directly causes Y in (130) while the instrument Z only causes Y though it impact on T.

$$\text{First Stage: } T = \kappa_1 + \beta_1 \cdot Z + \psi_1 \cdot X + \epsilon_1, \tag{122}$$

$$\text{Second Stage: } Y = \kappa_2 + \beta_2 \cdot T + \psi_2 \cdot X + \epsilon_2. \tag{123}$$

The 2SLS model (129)–(130) relies on the assumption that the instrument Z and control covariates X are independent of error term ϵ_2, that is, $(Z, X) \perp\!\!\!\perp \epsilon_2$. The 2SLS estimator $\hat{\beta}_2$ for parameter β_2 is expressed by Equation (131) and it is a consistent estimator under model assumptions.

$$\text{plim}(\hat{\beta}_2) = \frac{\text{cov}(Z, Y)\,\text{cov}(X, X) - \text{cov}(Y, X)\,\text{cov}(Z, X)}{\text{cov}(Z, T)\,\text{cov}(X, X) - \text{cov}(T, X)\,\text{cov}(Z, X)} = \beta_2. \tag{124}$$

The 2SLS estimator $\hat{\psi}_2$ for parameter ψ_2 is expressed by Equation (132) and it is a consistent estimator under model assumptions.

$$\text{plim}(\hat{\psi}_2) = -\frac{\text{cov}(Z, Y)\,\text{cov}(T, X) - \text{cov}(Y, X)\,\text{cov}(Z, T)}{\text{cov}(Z, T)\,\text{cov}(X, X) - \text{cov}(T, X)\,\text{cov}(Z, X)} = \psi_2. \tag{125}$$

Each of the identification formulas for the causal effects in (122)–(125) describes a ratio of covariances that corresponds to one of the three 2SLS formulas (128), (131) or (131).

The effect of choice T on mediator M is given by:

$$\frac{dE(M(t))}{dt} = \widetilde{\varphi}_T = \frac{\text{cov}(Z, M)}{\text{cov}(Z, T)}.$$

According to Equation (128), this effect can be estimated by the 2SLS regression (126)–(127) in which Z is the instrument, T is the endogenous variable and M is the outcome.

The total effect of T on outcome Y is given by:

$$\frac{dE(Y(t))}{dt} = \widetilde{\varphi}_T \cdot \beta_M + \widetilde{\beta}_T = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,T)}.$$

According to Equation (128),this effect can be estimated by the 2SLS regression (126)–(127) in which Z is the instrument, T is the endogenous variable and Y is the outcome.

The causal effect of mediator M on outcome Y is given by:

$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\text{cov}(Z,Y)\,\text{cov}(T,T) - \text{cov}(Y,T)\,\text{cov}(Z,T)}{\text{cov}(Z,M)\,\text{cov}(T,T) - \text{cov}(M,T)\,\text{cov}(Z,T)},$$

which can be estimated by the 2SLS regression (126)–(127) where Z is the instrument, T is the endogenous variable and M is the outcome.

The causal effect of mediator M on outcome M is given by:

$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\text{cov}(Z,Y)\,\text{cov}(T,T) - \text{cov}(Y,T)\,\text{cov}(Z,T)}{\text{cov}(Z,M)\,\text{cov}(T,T) - \text{cov}(M,T)\,\text{cov}(Z,T)}.$$

According to the 2SLS estimator in (131), this causal effect can be estimated by $\hat{\beta}_2$ in the 2SLS regression (129)–(130) in which Z plays the role of the instrument, M is the endogenous variable, T is the control covariate and Y is the outcome.

The Indirect Effect of choice T on outcome Y is given by:

$$\frac{\partial E(Y(t,m))}{\partial m} = \widetilde{\beta}_T = \frac{\text{cov}(Z,M)\,\text{cov}(T,Y) - \text{cov}(Z,Y)\,\text{cov}(T,M)}{\text{cov}(T,T)\,\text{cov}(Z,M) - \text{cov}(Z,T)\,\text{cov}(T,M)}.$$

According to the 2SLS estimator in (132), this causal effect can be estimated by $\hat{\psi}_2$ in the 2SLS regression (129)–(130) in which Z plays the role of the instrument, M is the endogenous variable, T is the control covariate and Y is the outcome.

Online Appendix J Total, Indirect and Direct Effects under One Instrument

Online Appendix H.2 describes a linear mediation model whose primary causal effects are identified by the following equations:

$$\text{Total Effect of } T \text{ on } Y : \frac{dE(Y(t))}{dt} = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,T)}. \tag{126}$$

$$\text{Direct Effect of } T \text{ on } Y : \frac{\partial E(Y(t,m))}{\partial t} = \frac{\text{cov}(Z,M)\,\text{cov}(T,Y) - \text{cov}(Z,Y)\,\text{cov}(T,M)}{\text{cov}(T,T)\,\text{cov}(Z,M) - \text{cov}(Z,T)\,\text{cov}(T,M)}. \tag{127}$$

$$\text{Effect of } M \text{ on } Y : \frac{\partial E(Y(t,m))}{\partial m} = \frac{\text{cov}(Z,T)\,\text{cov}(T,Y) - \text{cov}(T,T)\,\text{cov}(Z,Y)}{\text{cov}(T,M)\,\text{cov}(Z,T) - \text{cov}(T,T)\,\text{cov}(Z,M)} \tag{128}$$

$$\text{Effect of } T \text{ on } M : \frac{dE(M(t))}{dt} = \frac{\text{cov}(Z,M)}{\text{cov}(Z,T)} \tag{129}$$

$$\text{Indirect Effect of } T \text{ on } Y : \frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}. \tag{130}$$

The literature of mediation analysis typically expresses the total effect of T on Y as the sum of its direct and indirect effects. In our notation, this decomposition is is stated as following:

$$\underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}} = \underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}}. \tag{131}$$

We show that the decomposition described in (138) is exact in the case of a single instrument. That is to say that the covariance ratio that identifies the total effect of T on Y in equation (133) is equal to the covariance ratio that identifies the direct effect in Equations (134) plus the multiplication of the covariance ratios that identify the effect of T on M in (136) and the effect of M on Y

described in Equation (135). We thank David Slichter for pointing this fact.

$$
\begin{aligned}
&\underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}} \\
&= \underbrace{\frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}}_{\frac{\partial E(Y(t,m))}{\partial t}} + \underbrace{\frac{\operatorname{cov}(Z,T)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)}}_{\frac{\partial E(Y(t,m))}{\partial m}} \cdot \underbrace{\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}}_{\frac{dE(M(t))}{dt}} \\
&= \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} + \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)} \\
&= \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} + \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,M)\operatorname{cov}(T,Y)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)\frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \left(\frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}\right) \cdot \left(\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}\right) \\
&= \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} = \underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}}.
\end{aligned}
$$

The first equality expresses the total effect of T on Y in terms of its direct and indirect effects. The second equality substitutes the direct and indirect effects by their identification formulas described in (134), (135) and (133). The third equation isolates and eliminates the common term $\operatorname{cov}(Z,M)$ in the denominator of $\frac{dE(Y(m))}{dm}$. The fourth equation flips the sign of the terms in the last covariance ratio. Now the overall sum has the same denominator. The fifth equation eliminates the common term in the sum of the numerators of both ratios. The sixth equation exchange the covariances $\operatorname{cov}(Z,M)$ and $\operatorname{cov}(Z,Y)$ of the first term of the numerator. The seventh equation includes the term $\frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}$ which is equal to one. The eight equation exchange the covariances $\operatorname{cov}(Z,Y)$ and $\operatorname{cov}(Z,T)$ of the second term of the numerator. The night equation isolates the common denominator of the expression. The tenth equation eliminates the common first term of both numerator and denominator. The resulting formula is the covariate ratio $\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}$ which, according to (133), is equal to the total effect of choice T on outcome Y.

Online Appendix K Model Specification Test

The nonparametric version of the restricted model is given by the following equations:

$$T = f_T(Z, V_T, \epsilon_T), \tag{132}$$
$$U = f_U(T, \epsilon_U), \tag{133}$$
$$M = f_M(T, U, V_T, \epsilon_M), \tag{134}$$
$$Y = f_Y(T, M, U, V_Y, \epsilon_Y), \tag{135}$$
$$Z \perp\!\!\!\perp V_Y \perp\!\!\!\perp V_T \perp\!\!\!\perp \epsilon_Y \perp\!\!\!\perp \epsilon_M \perp\!\!\!\perp \epsilon_U \perp\!\!\!\perp \epsilon_T. \tag{136}$$

The identification of causal effects in our mediation model exploits three exclusion restrictions:

$$Z \perp\!\!\!\perp Y(t), \quad Z \perp\!\!\!\perp M(t), \quad \text{and} \quad Z \perp\!\!\!\perp Y(m)|T. \tag{137}$$

Exclusion Restrictions (144) do not depend on the linearity and hold for the nonparametric model defined by (139)–(143). Exclusion restrictions alone do not guarantee identification. The literature on instrumental variables offers a range of additional assumptions that enable identification of causal effects. Examples of such additional assumptions are: monotonicity (Imbens and Angrist, 1994), separability of (Heckman and Vytlacil, 2005), control function (Blundell and Powell, 2004) or revealed preference analysis (Pinto, 2015).

The three exclusion restrictions in (144) arise from the restrictions imposed on the causal relations of the unobserved variables V_T and V_Y. Specifically the Mediation Model (139)–(143) assumes that V_T jointly causes T and M while V_Y causes M and Y jointly, but neither V_T or V_Y causes T, M and Y simultaneously. A more general model allows for a common unobserved variable V causes T, M and Y simultaneously, as described by Equations (145)–(149).

$$T = f_T(Z, V, \epsilon_T), \tag{138}$$
$$U = f_U(T, \epsilon_U), \tag{139}$$
$$M = f_M(T, U, V, \epsilon_M), \tag{140}$$
$$Y = f_Y(T, M, U, V, \epsilon_Y), \tag{141}$$
$$Z \perp\!\!\!\perp V \perp\!\!\!\perp \epsilon_Y \perp\!\!\!\perp \epsilon_M \perp\!\!\!\perp \epsilon_U \perp\!\!\!\perp \epsilon_T. \tag{142}$$

Exclusion restriction $Z \perp\!\!\!\perp Y(m)|T$ does not hold in the general model described by Equations (145)–(149). Exclusion restrictions $Z \perp\!\!\!\perp Y(t)$ and $Z \perp\!\!\!\perp M(t)$ hold for both the Restricted Model (139)–(143) and the General Model (145)–(149).

For sake of clarity, we term the mediation model described by (139)–(143) as the Restricted Model. In contrast, we term the mediation model described by (145)–(149) as the General model.

Our goal is to test whether the Restricted Model (139)–(143) holds instead of the General model (145)–(149). That is to say, we want to test whether an unobserved random variable V jointly causes the treatment T, Mediator M and outcome Y. To do so, we evoke the linearity assumption of Online Appendix H that is used in the empirical estimation of our mediation model. Our aim is not to test those linear assumptions, instead we assume linearity to test the causal relations in the Restricted Model (139)–(143) against the ones in the General Model (145)–(149).

We show that an instrumental variable consisting of a single variable does not generate a test that infers if the Restricted Model (139)–(143) is rejected in favor of the General Model (145)–(149). nevertheless we show that a model specification test can be generated if we have two instrumental variables. By two instruments we mean two random variables that cause the treatment variable

T but are not caused by the unobserved confounding variables. As mentioned, we maintain the assumption of linearity for both General and Restricted models. Our test bares some some similarities with the the Sargan-Hansen test that exploits model over-identifying restrictions to do inference on model coefficients.

Online Appendix K.1 The Case of a General Model with a Single Instrumental Variable

In this section we compare the Restricted Mediation Model that is assumed to hold in our estimations with the General Mediation Model that allows for an unobserved variable V to cause the three main variables we examine: treatment T, Mediator M and outcome Y.

Figure 7 summarizes key properties and differences of these two models. Panel A of Figure 7 presents the Directed Acyclic Graphs (DAG) of the restricted model examined in the paper and the General Mediation model that does not assume the restriction on the causal restriction on confounding variables. Panel B presents the structural equations associated with each model. Panel C presents the linear equations that subsume the causal relations described in each model. Panel D displays the equalities generated by the covariance structure arising from the linear equations. Those are used to identify model coefficients.

Let the General Linear mediation Model with one Instrumental Variable be described by the following equations:

$$
\begin{aligned}
\text{Instrumental Variable } Z &= \epsilon_Z, && (143)\\
\text{Treatment } T &= \xi_Z \cdot Z + \xi_V \cdot V + \epsilon_T, && (144)\\
\text{Unobserved Mediator } U &= \zeta_T \cdot T + \epsilon_U, && (145)\\
\text{Observed Mediator } M &= \varphi_T \cdot T + \varphi_U \cdot U + \delta \cdot V + \epsilon_M, && (146)\\
\text{Outcome } Y &= \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V + \epsilon_Y, && (147)\\
\text{Exogenous Variables} \quad & Z, V, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y \text{ are statistically independent variables,} && (148)\\
\text{Scalar Coefficients} \quad & \xi_Z, \xi_V, \zeta_T, \varphi_T, \varphi_U, \delta, \beta_T, \beta_M, \beta_U, \beta_V && (149)\\
\text{Unobserved Variables} \quad & V, U, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y. && (150)
\end{aligned}
$$

We assume that all variables have mean zero. This assumption does not incur in less of generality, but simplify notation as intercepts can be suppressed.

The main difference between the Restricted Model (Equations (74)–(81) of Online Appendix H) and the (Equations (150)–(157) above) resides on the equations that define the data generating process of the mediator M. The Restricted Model evokes two terms $\delta_Y \cdot V_Y$ and $\delta_T \cdot V_T$ in Equation (153) while these terms are subsumed by the term $\delta \cdot V$ in Equation (153) of the General Model. The rest of the coefficients share the same notation of coefficients in both models. The unobserved variable V_Y that causes outcome Y in the Restricted Model (78) is replaced by the unobserved variable V in the General Model (154).

We eliminate the unobserved mediator U from Equations (153)–(154) in the same fashion that U is eliminated from Equations (77)–(78) of Online Appendix H. The new equations are:

$$
\begin{aligned}
M &= (\varphi_T + \varphi_U \zeta_T) \cdot T + \varphi_U \cdot \epsilon_U + \delta \cdot V + \epsilon_M, && (151)\\
Y &= (\beta_T + \beta_U \zeta_T) \cdot T + \beta_M \cdot M + \beta_U \cdot \epsilon_U + \beta_V \cdot V + \epsilon_Y. && (152)
\end{aligned}
$$

We use the same change in notation as performed in the restricted model: $\widetilde{\varphi}_T = \varphi_T + \varphi_U \zeta_T$,

$\widetilde{\beta}_T = \beta_T + \beta_U \zeta_T$, and $\widetilde{U} = \epsilon_U$. Equation (150)–(154) are therefore simplified to:

$$\begin{aligned}
\text{Instrumental Variable } Z &= \epsilon_Z, \tag{153}\\
\text{Treatment } T &= \xi_Z \cdot Z + \xi_V \cdot V + \epsilon_T, \tag{154}\\
\text{Observed Mediator } M &= \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta \cdot V + \epsilon_M, \tag{155}\\
\text{Outcome } Y &= \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V + \epsilon_Y. \tag{156}
\end{aligned}$$

Model (160)–(163) can be conveniently expressed in matrix notation: Equation (164) of the General Model is the counterpart of Equation (101) of the restricted model in Online Appendix H.2.

Following previous notation, we use $\mathbf{X} = [Z, T, M, Y]'$ for the vector of observed variables, and $\varepsilon = [\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y]'$ for the vector of exogenous error terms. The vector the vector of unobserved variables that generate endogenous variables is defined as $\mathbf{V}_G = [V, \widetilde{U}]'$. According to (155), the random vectors $\mathbf{V}_G$ and ε are independent, that is, $\mathbf{V}_G \perp\!\!\!\perp \varepsilon$. We use $\mathbf{K}$ for the matrix of parameters that multiply $\mathbf{X}$ and $\mathbf{A}_G$ for the matrix of parameters that multiply $\mathbf{V}_G$.

$$\mathbf{X} = \begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}, \quad \mathbf{A}_G = \begin{bmatrix} 0 & 0 \\ \xi_V & 0 \\ \delta & \varphi_U \\ \beta_V & \beta_U \end{bmatrix}, \quad \mathbf{V}_G = \begin{pmatrix} V \\ \widetilde{U} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}. \tag{157}$$

Using the notation defined in (101), we can express the linear system (87)–(90) as following:

$$\underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}}_{\mathbf{K}} \cdot \underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} 0 & 0 \\ \xi_V & 0 \\ \delta & \varphi_U \\ \beta_V & \beta_U \end{bmatrix}}_{\mathbf{A}_G} \cdot \underbrace{\begin{pmatrix} V \\ \widetilde{U} \end{pmatrix}}_{\mathbf{V}_G} + \underbrace{\begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}}_{\varepsilon}, \tag{158}$$

$$\mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A}_G \cdot \mathbf{V}_G + \varepsilon. \tag{159}$$

The identification of coefficients in $\mathbf{K}, \mathbf{A}_G$ depends on the covariance matrix of the observed data. We follow the notation of Online Appendix H.2 closely. We use $\Sigma_{\mathbf{X}} = \text{cov}(\mathbf{X}, \mathbf{X})$ for the covariance matrix of observed variables $\mathbf{X}$, and $\Sigma_\varepsilon = \text{cov}(\varepsilon, \varepsilon)$ for the vector of error terms ε. Σ_ε is a diagonal matrix due to statistical independence of error terms. We also use $\Sigma_{\mathbf{V}_G} = \text{cov}(\mathbf{V}_G, \mathbf{V}_G)$ for the covariance of unobserved variables $\mathbf{V}_G$. The unobserved variables in $\mathbf{V}_G$ are statistically independent and have variance one, thus we have that $\Sigma_{\mathbf{V}_G} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix of dimension 2. Moreover, $\mathbf{V}_G \perp\!\!\!\perp \varepsilon$ implies that $\text{cov}(\mathbf{V}_G, \varepsilon) = \mathbf{0}$, where $\mathbf{0}$ is a matrix of zero elements.

Equation (167) determines the relation between the covariance matrices of observed and unobserved variables:

$$\mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A}_G \cdot \mathbf{V}_G + \varepsilon \Rightarrow (\mathbf{K} - \mathbf{I})\, \Sigma_{\mathbf{X}}\, (\mathbf{K} - \mathbf{I})' = \mathbf{A}_G \mathbf{A}_G' + \Sigma_\varepsilon, \tag{160}$$

where the second equation is due to $\mathbf{V}_G \perp\!\!\!\perp \varepsilon$ and the third equations comes from $\Sigma_{\mathbf{V}_G} = \mathbf{I}$. Equation (167) generates ten equalities: four equalities are generated by the equality of the diagonal of the covariance matrices and six equations from the off-diagonal relation of the covariance matrices.

The identification analysis of the coefficients of the General Model arises from the six off-diagonal equations generated by (167). Those are listed below:

$$\mathrm{cov}(Z,T) - \mathrm{cov}(Z,Z) \cdot \xi_Z = 0 \tag{161}$$

$$\mathrm{cov}(Z,M) - \mathrm{cov}(Z,T) \cdot \widetilde{\varphi}_T = 0 \tag{162}$$

$$\mathrm{cov}(Z,Y) - \mathrm{cov}(Z,M) \cdot \beta_M - \mathrm{cov}(Z,T) \cdot \widetilde{\beta}_T = 0 \tag{163}$$

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,T) \cdot \widetilde{\beta}_T - \mathrm{cov}(T,M) \cdot \beta_M = \beta_V \cdot \xi_V \tag{164}$$

$$\mathrm{cov}(M,Y) - \mathrm{cov}(T,M) \cdot \widetilde{\beta}_T - \mathrm{cov}(M,M) \cdot \beta_M = \beta_U \cdot \varphi_U + \beta_V \cdot (\delta + \xi_V \cdot \widetilde{\varphi}_T) \tag{165}$$

$$\mathrm{cov}(T,M) - \mathrm{cov}(T,T) \cdot \widetilde{\varphi}_T = \delta \cdot \xi_V \tag{166}$$

Equations (168)–(173) of the General Mediation Model are the symmetric to Equations (107)–(112) of the Restricted Mediation Model.

Equations (168)–(169) are identical to Equations (107)–(108) of the restricted model. Therefore ξ_Z and $\widetilde{\varphi}_T$ are identified by the same covariance ratios presented in Equations (113)–(114).

Even though Equation (170) is identical to Equation (109) of the restricted model, Equation (171) differs from Equation (110). Equation (110) of the Restricted Model states that

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,T) \cdot \widetilde{\beta}_T - \mathrm{cov}(T,M) \cdot \beta_M = 0.$$

The counterpart of (171) in the General model states that

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,T) \cdot \widetilde{\beta}_T - \mathrm{cov}(T,M) \cdot \beta_M = \beta_V \cdot \xi_V.$$

As a consequence, $\widetilde{\beta}_T$ and β_M are not identified in the General Mediation Model. Moreover, the comparison of the covariance structure of both models does not allow to distinguish one model from the other. We conclude that a single instrument is insufficient to generate overidentifying restrictions that enable us to verify if observed data arises from the restricted or the general model.

Online Appendix K.2 Restricted Model with Multiple Instruments

In this section we show that two (or more) instrumental variables generate over-identifying restrictions that enables to perform a model specification test. The equations presented in this section follow closely the ones presented in Online Appendix H.2. The restricted model with two instrumental variables is described by the following equations:

$$\text{Vector of Instrumental Variables } \mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}, \tag{167}$$

$$\text{Treatment } T = \boldsymbol{\xi}_Z' \cdot \mathbf{Z} + \xi_V \cdot V_T + \epsilon_T, \text{ where } \boldsymbol{\xi}_Z = \begin{bmatrix} \xi_{Z,1} \\ \xi_{Z,2} \end{bmatrix}, \tag{168}$$

$$\text{Observed Mediator } M = \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{169}$$

$$\text{Outcome } Y = \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \tag{170}$$

Model (174)–(177) can be conveniently expressed in matrix notation in the same fashion as Equations (101) of Online Appendix H.2.

The covariance equation (106) of Online Appendix H.2 also holds. The identification of model coefficients relies on the equations governing the covariance matrix of observed variables. These identifying equations for the model (107)–(112) are given by:

$$\mathrm{cov}(\mathbf{Z},T) - \mathrm{cov}(\mathbf{Z},\mathbf{Z}) \cdot \boldsymbol{\xi}_Z = 0 \tag{171}$$

$$\mathrm{cov}(\mathbf{Z},M) - \mathrm{cov}(\mathbf{Z},T) \cdot \widetilde{\varphi}_T = 0 \tag{172}$$

$$\mathrm{cov}(\mathbf{Z},Y) - \mathrm{cov}(\mathbf{Z},M) \cdot \beta_M - \mathrm{cov}(\mathbf{Z},T) \cdot \widetilde{\beta}_T = 0 \tag{173}$$

$$\mathrm{cov}(T,Y) - \mathrm{cov}(T,T) \cdot \widetilde{\beta}_T - \mathrm{cov}(T,M) \cdot \beta_M = 0 \tag{174}$$

$$\mathrm{cov}(M,Y) - \mathrm{cov}(T,M) \cdot \widetilde{\beta}_T - \mathrm{cov}(M,M) \cdot \beta_M = \beta_U \cdot \varphi_U + \beta_V \cdot \delta_Y \tag{175}$$

$$\mathrm{cov}(T,M) - \mathrm{cov}(T,T) \cdot \widetilde{\varphi}_T = \delta_T \cdot \xi_V \tag{176}$$

Equation (178) represent a system of two linear equations associated with each instrumental variable in $\mathbf{Z} = [Z_1, Z_2]'$. Equation (178) enables the identification of the vector of coefficients $\boldsymbol{\xi}_Z = [\xi_{Z,1}, \xi_{Z,2}]'$.

Equation (179) also represents a system of two linear equations that allows for the identification of the coefficient $\widetilde{\varphi}_T$. Parameter $\widetilde{\varphi}_T$ is overidentified as there are two linear equations that allow for the identification of the parameter.

Equation (180) represents a system of two linear equations that enable us to identify two coefficients: β_M and $\widetilde{\beta}_T$. Equation (181) constitute an overidentification restriction for parameters β_M and $\widetilde{\beta}_T$. This result differs from the identification using a single instrumental variable. We explain in Online Appendix H.2 that parameters β_M and $\widetilde{\beta}_T$ required the use of two equations, that is (109) and (110), which are the counterpart of equations (180) and (181) above. In summary, Equation (181) constitute an overidentified restriction in the case of two instrumental variables while it is a necessary equation for the identification of $\beta_M, \widetilde{\beta}_T$ when the dimension of the instrumental variable is equal to one.

Equations (181)–(183) are identical to (110)–(112) in Online Appendix H.2. Equation (182) enables the identification of the sum $\beta_U \cdot \varphi_U + \beta_V \cdot \delta_Y$ and Equation (183) identifies $\delta_T \cdot \xi_V$.

We conclude that the advent of more than one instrument does not render the identification of additional parameters. Instead it changes the identification status of parameters $\widetilde{\varphi}_T$, β_M and $\widetilde{\beta}_T$ from just-identified to overidentified.

Online Appendix K.3 General Model with Multiple Instruments

The general model allows for an unobserved variable V to cause T, M and Y jointly. The equations presented in this section follow the ones presented in Online Appendix K.1. The general model with two instrumental variables stems from Equations (160)–(163) of Online Appendix K.1. The equations of the general model with two instrumental variables are displayed below:

$$\text{Vector of Instrumental Variables } \mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}, \tag{177}$$

$$\text{Treatment } T = \boldsymbol{\xi}'_Z \cdot \mathbf{Z} + \xi_V \cdot V + \epsilon_T, \text{ where } \boldsymbol{\xi}_Z = \begin{bmatrix} \xi_{Z,1} \\ \xi_{Z,2} \end{bmatrix}, \tag{178}$$

$$\text{Observed Mediator } M = \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta \cdot V + \epsilon_M, \tag{179}$$

$$\text{Outcome } Y = \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V + \epsilon_Y. \tag{180}$$

We can retrace the same steps described in Online Appendix K.1 to generate the following identifying equations:

$$\text{cov}(\mathbf{Z},T) - \text{cov}(\mathbf{Z},\mathbf{Z}) \cdot \boldsymbol{\xi}_Z = 0 \tag{181}$$

$$\text{cov}(\mathbf{Z},M) - \text{cov}(\mathbf{Z},T) \cdot \widetilde{\varphi}_T = 0 \tag{182}$$

$$\text{cov}(\mathbf{Z},Y) - \text{cov}(\mathbf{Z},M) \cdot \beta_M - \text{cov}(\mathbf{Z},T) \cdot \widetilde{\beta}_T = 0 \tag{183}$$

$$\text{cov}(T,Y) - \text{cov}(T,T) \cdot \widetilde{\beta}_T - \text{cov}(T,M) \cdot \beta_M = \beta_V \cdot \xi_V \tag{184}$$

$$\text{cov}(M,Y) - \text{cov}(T,M) \cdot \widetilde{\beta}_T - \text{cov}(M,M) \cdot \beta_M = \beta_U \cdot \varphi_U + \beta_V \cdot (\delta + \xi_V \cdot \widetilde{\varphi}_T) \tag{185}$$

$$\text{cov}(T,M) - \text{cov}(T,T) \cdot \widetilde{\varphi}_T = \delta \cdot \xi_V \tag{186}$$

Equations (188)–(190) of the general model with two instrumental variables are identical to Equations (178)–(180) of the restricted model. Those equations enable the identification of the parameters $\boldsymbol{\xi}_Z, \widetilde{\varphi}_T, \beta_M, \widetilde{\beta}_T$. Equations (192)–(193) of the general model are also identical to Equations (182)–(183) of the restricted model.

The key difference between the two models arise from Equation (181) of the Restricted Model and Equation (191) of the General Model. While (181) states that $\text{cov}(T,Y) - \text{cov}(T,T) \cdot \widetilde{\beta}_T - \text{cov}(T,M) \cdot \beta_M = 0$, Equation (191) states that $\text{cov}(T,Y) - \text{cov}(T,T) \cdot \widetilde{\beta}_T - \text{cov}(T,M) \cdot \beta_M = \beta_V \cdot \xi_V$.

Online Appendix K.4 Inference on General versus the Restricted Models

A simple model specification test can be performed by exploring Equations (180)–(181) of the Restricted model. Those two equations can be interpreted as a consequence of two moment conditions presented as following:

$$E(\mathbf{Z} \cdot (Y - \widetilde{\beta}_T \cdot T - \beta_M \cdot M)) = 0 \quad \Rightarrow \quad \text{cov}(\mathbf{Z},Y) = \text{cov}(\mathbf{Z},M) \cdot \beta_M + \text{cov}(\mathbf{Z},T) \cdot \widetilde{\beta}_T \tag{187}$$

$$E(T \cdot (Y - \widetilde{\beta}_T \cdot T - \beta_M \cdot M)) = 0 \quad \Rightarrow \quad \text{cov}(T,Y) = \text{cov}(T,T) \cdot \widetilde{\beta}_T + \text{cov}(T,M) \cdot \beta_M \tag{188}$$

Moment Conditions (194)–(195) can be combined into a single equality using the matrix notation below:

$$E\left(\begin{bmatrix}\mathbf{Z}\\ T\end{bmatrix} \cdot \left(Y - [M\ ,\ T] \cdot \begin{bmatrix}\beta_M\\ \widetilde{\beta}_T\end{bmatrix}\right)\right) = 0 \quad \Rightarrow \quad \begin{bmatrix}\text{cov}(\mathbf{Z},Y)\\ \text{cov}(T,Y)\end{bmatrix} = \begin{bmatrix}\text{cov}(\mathbf{Z},M) & \text{cov}(\mathbf{Z},T)\\ \text{cov}(T,M) & \text{cov}(T,T)\end{bmatrix} \cdot \begin{bmatrix}\beta_M\\ \widetilde{\beta}_T\end{bmatrix} \tag{189}$$

We can apply the Generalized Method of Moments (GMM) of Hansen (1982) to the Moment Condition (196). The GMM estimator generated by moment (196) is given by:

$$\begin{bmatrix}\widehat{\beta_{M,1}}\\ \widehat{\widetilde{\beta}_{T,1}}\end{bmatrix} = \left([\text{M},\text{T}]' \cdot \mathbf{P}_{\text{Z,T}}\,[\text{M},\text{T}]\right)^{-1} \cdot \left([\text{M},\text{T}]' \cdot \mathbf{P}_{\text{Z,T}} \cdot \text{Y}\right), \tag{190}$$

$$\text{such that } \mathbf{P}_{\text{Z,T}} = [\text{Z},\text{T}] \cdot \left([\text{Z},\text{T}]' \cdot [\text{Z},\text{T}]\right)^{-1} \cdot [\text{Z},\text{T}]', \tag{191}$$

where $\text{T},\text{M},\text{Y}$ are $N \times 1$ data vectors associated respectively with observed variables T, M, Y; Z is a $N \times K$ matrix of data associated with K instrumental variables; N denotes sample size and $\mathbf{P}_{\text{Z,T}}$ stands for the orthogonal projection on the space generated by the columns of $[\text{Z},\text{T}]$.

The GMM estimator in (197) can be interpreted as a Two Stage Least Square regression (129)–(130), in which $\mathbf{Z}$ plays the role of instrumental variables, M is the endogenous variable, T is a conditioning variable in both first and second stages and Y is the outcome.

The General Mediation Model (184)–(187) differs from the restricted model as Equality (195) does not hold. Nevertheless, the GMM method can be applied to Moment Condition (194), that is, $E(\mathbf{Z} \cdot (Y - \widetilde{\beta}_T \cdot T - \beta_M \cdot M)) = 0$, which still hold in the general model. GMM estimator (197) for the the general model that is based only on Moment Condition (194) is given by:

$$\begin{bmatrix} \widehat{\beta_{M,2}} \\ \widehat{\widetilde{\beta}_{T,2}} \end{bmatrix} = \left([\mathrm{M}, \mathrm{T}]' \cdot \mathbf{P}_{\mathrm{Z}} \, [\mathrm{M}, \mathrm{T}]\right)^{-1} \cdot \left([\mathrm{M}, \mathrm{T}]' \cdot \mathbf{P}_{\mathrm{Z}} \cdot \mathrm{Y}\right), \tag{192}$$

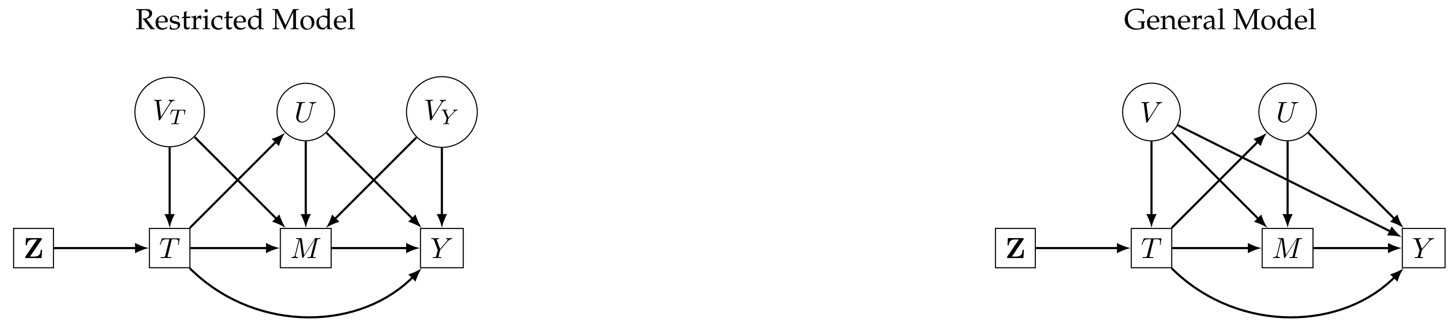
$$\text{such that } \mathbf{P}_{\mathrm{Z}} = \mathrm{Z} \cdot \left(\mathrm{Z}' \cdot \mathrm{Z}\right)^{-1} \cdot \mathrm{Z}'. \tag{193}$$

The GMM Estimator (199) can be interpreted as the Two Stage Least Square regression (126) (127) in which $\mathbf{Z}$ plays the role of instrumental variables, Y is the outcome and both M and T are endogenous variables.

If the causal assumptions of the restricted model hold, then both GMM estimators (197) and (199) provide consistent estimates of β_M and $\widetilde{\beta}_T$. If the causal assumptions of the restricted model do not hold, then (197) does not generates a consistent estimate of $\widehat{\beta_M}, \widehat{\widetilde{\beta}_T}$. Thus large differences between the estimates $\widehat{\beta_{M,1}}, \widehat{\widetilde{\beta}_{T,1}}$ of (197) versus $\widehat{\beta_{M,2}}, \widehat{\widetilde{\beta}_{T,2}}$ of (199) provide statistical evidence gainst the null hypothesis that the causal assumptions of the Restrictive Model holds.

Online Appendix Table 7: Restricted and General Mediation Model with One Instrumental Variable

A. DAG Representation



B. Structural Equations

$T = f_T(\mathbf{Z}, V_T, \epsilon_T)$	$T = f_T(\mathbf{Z}, V, \epsilon_T)$
$U = f_U(T, \epsilon_U)$	$U = f_U(T, \epsilon_U)$
$M = f_M(T, U, V_T, \epsilon_M)$	$M = f_M(T, U, V, \epsilon_M)$
$Y = f_Y(T, M, U, V_Y, \epsilon_Y)$	$Y = f_Y(T, M, U, V_Y, \epsilon_Y)$
$\mathbf{Z} \perp\!\!\!\perp V_Y \perp\!\!\!\perp V_T \perp\!\!\!\perp \epsilon_Y \perp\!\!\!\perp \epsilon_M \perp\!\!\!\perp \epsilon_U \perp\!\!\!\perp \epsilon_T$	$\mathbf{Z} \perp\!\!\!\perp V \perp\!\!\!\perp \epsilon_Y \perp\!\!\!\perp \epsilon_M \perp\!\!\!\perp \epsilon_U \perp\!\!\!\perp \epsilon_T$

C. Linear Equations

$T = \boldsymbol{\xi}_Z \cdot \mathbf{Z} + \xi_V \cdot V_T + \epsilon_T$	$T = \boldsymbol{\xi}_Z \cdot \mathbf{Z} + \xi_V \cdot V + \epsilon_T$
$M = \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M$	$M = \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta \cdot V + \epsilon_M$
$Y = \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y$	$Y = \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V + \epsilon_Y$

D. Identifying Equations

$\text{cov}(\mathbf{Z}, T) - \text{cov}(\mathbf{Z}, \mathbf{Z}) \cdot \boldsymbol{\xi}_Z = 0$	$\text{cov}(\mathbf{Z}, T) - \text{cov}(\mathbf{Z}, \mathbf{Z}) \cdot \boldsymbol{\xi}_Z = 0$
$\text{cov}(\mathbf{Z}, M) - \text{cov}(\mathbf{Z}, T) \cdot \widetilde{\varphi}_T = 0$	$\text{cov}(\mathbf{Z}, M) - \text{cov}(\mathbf{Z}, T) \cdot \widetilde{\varphi}_T = 0$
$\text{cov}(\mathbf{Z}, Y) - \text{cov}(\mathbf{Z}, M) \cdot \beta_M - \text{cov}(\mathbf{Z}, T) \cdot \widetilde{\beta}_T = 0$	$\text{cov}(\mathbf{Z}, Y) - \text{cov}(\mathbf{Z}, M) \cdot \beta_M - \text{cov}(\mathbf{Z}, T) \cdot \widetilde{\beta}_T = 0$
$\text{cov}(T, Y) - \text{cov}(T, M) \cdot \beta_M - \text{cov}(T, T) \cdot \widetilde{\beta}_T = 0$	$\text{cov}(T, Y) - \text{cov}(T, M) \cdot \beta_M - \text{cov}(T, T) \cdot \widetilde{\beta}_T = \beta_V \cdot \xi_V$
$\text{cov}(M, Y) - \text{cov}(M, M) \cdot \beta_M - \text{cov}(T, M) \cdot \widetilde{\beta}_T = \beta_U \cdot \varphi_U + \beta_V \cdot \delta_Y$	$\text{cov}(M, Y) - \text{cov}(M, M) \cdot \beta_M - \text{cov}(T, M) \cdot \widetilde{\beta}_T = \beta_U \cdot \varphi_U + \beta_V \cdot (\delta + \xi_V \cdot \widetilde{\varphi}_T)$
$\text{cov}(T, M) - \text{cov}(T, T) \cdot \widetilde{\varphi}_T = \delta_T \cdot \xi_V$	$\text{cov}(T, M) - \text{cov}(T, T) \cdot \widetilde{\varphi}_T = \delta \cdot \xi_V$

Panel A presents the Directed Acyclic Graphs (DAG) of the restricted model examined in the paper and the General Mediation model that does not assume the restriction on the causal restriction on confounding variables. Panel B presents the structural equations associated with each model. Panel C presents the linear equations that subsume the causal relations described in each model. Panel D displays the equalities generated by the covariance structure arising from the linear equations.