I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14293

Revisiting Gender Identity and Relative Income within Households: A Cautionary Tale on the Potential Pitfalls of Density Estimators

Daniel Kuehnle
Michael Oberfichtner
Kerstin Ostermann

APRIL 2021

DISCUSSION PAPER SERIES

IZA DP No. 14293

Revisiting Gender Identity and Relative Income within Households: A Cautionary Tale on the Potential Pitfalls of Density Estimators

Daniel Kuehnle
University of Duisburg-Essen and CINCH and IZA

Kerstin Ostermann
IAB

Michael Oberfichtner
IAB

APRIL 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Revisiting Gender Identity and Relative Income within Households: A Cautionary Tale on the Potential Pitfalls of Density Estimators*

We show that Bertrand et al.'s (QJE 2015) finding of a sharp drop in the relative income distribution within married couples at the point where wives start to earn more than their husbands is unstable across different estimation procedures and varies across contexts. We apply the estimators by McCrary (JoE, 2008, McC) and Cattaneo et al. (JASA, 2020, CJM) to administrative data from the US and Germany and compare their performance in a simulation. Large bins cause McC to substantially overreject the null hypothesis, and mass points close to the potential discontinuity affect McC more than CJM.

JEL Classification: C14, C18, D10, J16

Keywords: gender norms, relative income distribution, density estimation, US, Germany, replication

Corresponding author:
Daniel Kuehnle
University of Duisburg-Essen
Forsthausweg 2
47057 Duisburg
Germany

E-mail: daniel.kuehnle@uni-due.de

* This research was made possible through the use of Cornell University's Synthetic Data Server, which is funded through NSF Grant SES-1042181 and BCS-0941226. We especially thank Lori Beth Zehr, Veronica Roth and Jordan Stanley of the US Census Bureau who provided excellent support in using the SIPP/SSA/IRS Public Use Files. We thank Jessica Pan for sharing her Stata code with us. Moreover, we thank three anonymous referees, Jonas Jessen, Hendrik Schmitz and the conference participants at the 2019 meeting of the European Society of Population Economics, the 2019 meeting of Verein fuer Socialpolitik, and the 2019 meeting of the International Association of Applied Econometrics for helpful comments. Our final thanks go to the research data center at IAB who provided us with the couples data set for Germany. All remaining errors are our own.

1. Introduction

In a seminal paper, Bertrand et al. (2015, BKP henceforth) investigate income shares of married spouses in the US and report a discontinuity in the relative income distribution within couples at the point where wives start to earn more than their husbands. As standard economic theory (e.g., Becker, 1973) fails to predict such a drop, BKP interpret this discontinuity as evidence that wives avoid earning more than their husbands to comply with gender identity norms. Several papers have revisited the distribution of relative income within couples for different countries using the same approach.[1] The sharp drop in the relative income distribution also attracted substantial attention outside of academia (e.g., The Economist, 2012; NYT, 2018).

Somewhat surprisingly, BKP report only results using the McCrary (2008, McC henceforth) density test with data aggregated into 5% bins. Although this approach is compelling due to its simplicity and visual clarity, it could be problematic for two main reasons. First, 5% bins collapse the underlying distribution into 20 data points, which may oversmooth and hide important information from the data. Second, in a small share of couples, both partners earn identical incomes, which may arise for institutional or behavioural reasons, such as collective bargaining and tax minimization. As shown by Binder and Lam (fc.), Hederos Eriksson and Stenberg (2015), and Zinovyeva and Tverdostup (2018), the handling of such couples can substantially influence the results. These couples, though acknowledged by BKP, are however easily missed in a figure with 5% bins and should emerge more clearly with smaller bins.

Our paper makes three contributions. First, we demonstrate that large bins cause the McC test to reject the null hypothesis of no discontinuity too often. We complement earlier studies reporting that couples with equal income drive the discontinuity estimate by identifying this additional source of spurious discontinuities. Second, we contribute new evidence on the relative income distribution within couples accounting for couples with equal earnings. In particular, we addi-

[1] Using administrative income data from the US, Sweden, and Finland, Binder and Lam (fc.), Hederos Eriksson and Stenberg (2015), and Zinovyeva and Tverdostup (2018) find that such a discontinuity exists and explore alternative explanations for the observed patterns. Two studies use survey data (Codazzi et al., 2018; Sprengholz et al., 2019), but Murray-Close and Heggeness (2018) and Roth and Slotwinski (2018) provide evidence that couples around this threshold systematically misreport income in surveys; thus, administrative data appears superior to survey data in this context.

tionally apply the Cattaneo, Jansson, and Ma (2020, CJM henceforth) estimator to the US data, and we provide first evidence for Germany based on administrative income data using both estimators. Third, we document that the McC estimator reacts more strongly to the existence of a mass point close to the potential discontinuity than the CJM estimator.

Our analysis starts with a successful replication of BKP's original main finding.[2] We then examine how narrowing the bin size affects the results of the McC estimator. The intuition of the McC estimator, which tests for a discontinuity in the density function of the running variable at the cutoff, is fairly simple. First, the data needs to be finely pre-binned ("very undersmoothed", see McC p. 702) to construct a histogram of the running variable. Second, McC uses local linear regressions separately on each side of the cutoff to smooth out the histogram. Finally, McC tests for a discontinuity in the log of the estimated densities between both sides of the cutoff. The approach requires choosing two tuning parameters: the bin size for the histogram and the bandwidth for the regression.

Our paper focuses on the bin size as BKP deviate from McC's recommended procedure by manually setting the bin size to 5%, while following McC's suggestion for bandwidth selection. We use smaller bins to address the concerns about oversmoothing also applying the default bandwidth selection. The choice of bins size involves an inherent trade-off as smaller bin sizes allow for a finer visual inspection while potentially increasing any noise in the data. To balance this trade-off, we choose 1% bins as these should be less prone to oversmoothing but still yield estimates smooth enough for a clear graphical representation of the whole distribution.[3]

In the second step, we apply the nonparametric density estimator by Cattaneo, Jansson, and Ma (2020, CJM henceforth), which was not available at the time of the original BKP study. Just like McC, CJM test for discontinuities in density functions at specified cutoff values. The approach first obtains the variable's empirical distribution function and then estimate the density function

[2]We replicate only one small, though probably the most prominent, part of BKP. They also examine how relative income affects marriages, the wife's labour force participation and income, and home production.

[3]Again, this is an arbitrary choice and we could have picked even smaller bins, say 0.1%. We address this point by also using the McC default bins, which are on the order of 0.17% for the US. Although smaller bins may help to better detect a mass of couples with exactly equal earnings, smaller bins may also magnify the impact of this mass point on the estimated discontinuity.

as the first derivative of the distribution function on both sides of the cutoff. The estimator then tests for a discontinuity in the estimated density of the running variable at the specified cutoff. The CJM estimator is attractive because it avoids pre-binning the data and thereby uses all information from the underlying distribution. Moreover, it requires just one tuning parameter, the bandwidth. For consistent estimation, both estimators require standard regularity assumption and that the density function has several derivatives everywhere but at the cutoff. Given that both estimators provide consistent estimates of the size of the discontinuity under similar assumptions, we would expect them to yield similar conclusions if the assumptions hold. Crucially, the mass point of couples with identical incomes violates the smoothness assumption underlying both estimators. To satisfy this assumption, we also exclude the mass point and compare the results from the two estimators.

In the third step, we perform the same analysis using administrative income data on married couples from Germany, which is less gender egalitarian than the US.[4] In the presence of similar gender identity norms, we would expect the discontinuity to be at least as pronounced in Germany as in the US, if not larger. An interesting feature of this setting is that East and West Germany differ substantially in terms of social norms and female labour market attachment (e.g., Schnabel, 2016), which is likely to affect the relative income within couples; we therefore split the sample into East and West Germany.

In the fourth step, we explore potential reasons for different results across McC test specifications as well as between the McC and the CJM estimator. We perform a simulation exercise that compares rejection rates in absence of a discontinuity for different data generating processes (DGP) that closely resemble the distributions we observe in these settings.

Our replication using US data is related to Binder and Lam (fc.) who examine the same discontinuity from the perspective of a theoretical assortative matching model. They conclude that the inclusion of couples with equal incomes drives the discontinuity estimate. In contrast

[4]According to the Gender Social Norms Index provided by the United Nations (UNDP, 2020), Germans report less gender egalitarian norms with respect to education, economic outcomes, and physical integrity. Interestingly, this pattern holds if looking at men and women separately. Moreover, Germany has become less egalitarian in these domains between 2005 and 2014, whereas the US have become more egalitarian over the same time.

to our paper, they always use small bins and focus more on the role of bandwidth selection; their findings show that decreasing the bandwidth increases the influence of the mass point on the discontinuity estimate. We add to and complement their analysis by focusing on the step from 5% bins to smaller bins, i.e., towards McC's (p. 699) "finely gridded histogram" and by applying an alternative density estimator.

2. Replication of BKP for the US

We first provide a replication of the discontinuity in the relative income distribution using the Survey of Income and Program Participation (SIPP). We focus on the SIPP because it is linked to administrative income data which avoids systematical income misreporting of couples in surveys (for Switzerland, see Roth and Slotwinski, 2018). BKP and our study use the SIPP years 1990 through 2004. BKP include the first panel observation of married couples in which both partners earn a positive income and are 18 to 65 years old; we restrict the sample accordingly.[5]

We present the graphical results from our replication analysis in Figure 1. Panel A displays BKP's original graph and Panels B presents our replication results. Comparing the two panels shows that we can successfully replicate BKP's graph.[6] The McC estimator yields that the density function drops by 11.9% where wives start to outearn their husbands (i.e., at 0.500001, as BKP did), which is very close to the 12.3% reported by BKP (see Table 1, first column).

To investigate whether potential oversmoothing of the data affects the results, we next reduce the bin size to 1% and plot the estimated densities in Figure 2. Compared to 5% bins (Panel A), the finer 1% bins (Panel B) reveal more information from the underlying distribution and hint at the mass point right at the point of the supposed discontinuity. The McC test yields a discontinuity

[5]Our final sample size (69,500) differs slightly from the one of BKP (73,654). This difference is likely caused by different data access rules and reporting guidelines for external researchers. In contrast to BKP, we have to average and round all statistics over the four SIPP Gold Standard Files to comply with the disclosure guidelines of the US Census Bureau. For more information about the data and how to access it visit: `https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html`. Last accessed on March 11th, 2021.

[6]We use the Stata command *DCdensity* and slightly augment it to account for the fact that the wive's income share always lies between 0 and 1. These minor changes are necessary to obtain an accurate graphical representation of the density regarding wife's income shares close to 0 and 1 in all figures, but irrelevant for the discontinuity estimates as we get the same discontinuity estimates at 0.5 using both versions of *DCdensity*. For technical details, see Online Appendix A.1.

estimate of -14%. Using the McC default bins of size 0.17%, we obtain a discontinuity estimate of -15.3%. Thus, all McC specifications point towards a substantial discontinuity.

Next, we avoid pre-binning the data by using the fully automatic CJM estimator.[7] The discontinuity now turns insignificant and reduces to -6.7%. The estimated densities (Figure 2, Panel C) provide little support for a discontinuity; note that this figure also does not reveal the mass point at 0.5. That the two estimators yield different results raises concerns about the finding's robustness.

To reconcile the findings, we exclude the mass point in a final step. The discontinuity decreases substantially for the three specifications of the McC test (see Table 1), where the smaller bins indicate a drop of -6.5%, on the margin of statistical significance.[8] The CJM density test yields an insignificant estimate of -3.9%. Thus, once the assumptions of both estimators are likely fulfilled and the data is not too coarsely binned, we reach similar estimates, which indicate a small, if any, discontinuity in relative income in the US. We examine potential sources of differences between the estimators in Section 4.

3. Evidence for Germany

We now turn to Germany and follow the same methodological steps. We use administrative register data and focus on married couples in which both partners are employed subject to social security; our sampling restrictions mirror those of BKP.[9]

We begin with West Germany and present the estimated densities of relative income in Panels D to F in Figure 2. Similar to the finding from the US, using 5% bins reveals a substantial drop in the distribution of relative incomes. The McC test estimates a drop in the density of

[7] We use the Stata command *rddensity*, version 1.0, see Cattaneo et al. (2018) for details. An important programming difference between *rddensity* and *DCdensity* concerns how they treat observations that lie exactly at the supposed cutoff. Whereas *rddensity* treats them as being on the left of the cutoff, *DCdensity* treats them as being on the right of the cutoff. That said, this difference does not affect our results because we always test for a discontinuity just to the right of the cutoff, i.e., at 0.500001, as did BKP. Thereby, both commands treat the mass point as being on the left-hand side of the cutoff. We thank one anonymous referee for bringing this to our attention.

[8] Note that our estimates differ slightly from Binder and Lam (fc.). Excluding the mass point and using the default specification of McC, they find a drop of -3.4%, which is clearly insignificant. Whereas we restrict the sample to the years 1990 through 2004 which BKP used, Binder and Lam (fc.) additionally include the 1984 and 2008 SIPP data. These different samples likely cause the small difference in the estimations.

[9] See Online Appendix A.2 for details about the data source and sampling procedures.

0.369 log points (Table 1, Panel C). However, using 1% bins as in Panel E indicates a much smaller discontinuity, and the McC test estimates a discontinuity of -7.5%. The default bins produce similar estimates (-8.7%). Similar to our findings for the US, using the CJM estimator results in a smaller discontinuity estimate (-4.9%). The results are thus highly volatile across specifications when including couples in which both partners earn identical incomes.

Table 1 further shows that dropping all couples with identical incomes does not change the results from the coarsely binned data. However, the McC test using smaller bins and the CJM estimator yield very similar estimates ranging from -3.3% to -4.2%, all of which are statistically insignificant. These estimates do not support the existence of a discontinuity in West Germany and suggest that too coarsely binned data may yield misleading results.

Finally, we examine East Germany. Panels G to I in Figure 2 indicate that wives' earnings shares in East Germany follow an inverted V-shape – supportive of more gender egalitarian norms compared to West Germany. When including couples with equal incomes, all estimators yield similar results. The McC tests all indicate a substantial and significant discontinuity on the order of -20% (see Table 1), whereas the CJM estimator estimates a slightly smaller discontinuity (-15.7%). We again drop all couples with equal incomes. We still find consistent evidence for a discontinuity, ranging from -10.3% to -14.3%, though it becomes statistically insignificant for the CJM estimator.

4. Exploring the differences within and between estimators

As the results differ substantially across different McC specifications and between the two estimators, this section explores potential sources of the observed differences. For this purpose, we simulate three different DGPs that match the empirical distributions for the US, West and East Germany, but do not exhibit a discontinuity at 0.5, see Online Appendix A.3 for details. We then evaluate how differently sized, but still rather minor, mass points just to the left of the cut-off affect the performance of each estimator. Table 2 presents the rejection rates of the true null hypothesis (no discontinuity) at the 5%-level based on 2,000 replications for each specification.

Absent a mass point, the default version of the McC and CJM density tests both perform close to the nominal level of 5% in all three settings. However, using 5% bins increases McC's rejection rates above 90% in the settings that mirror the US and West Germany. Only in case of East Germany, the rejection rate is not affected by these large bins. Too large bins hence endanger the validity of the McC estimator.

Turning to the mass point of couples with equal incomes, we begin with a closer visual inspection and focus on couples in which the wives' income share lies between 0.45 and 0.55. Figure 3 plots the German data in small bins (0.1%) and overlays the estimated densities from the McC and CJM density tests.[10] Whereas the density estimated by the CJM estimator is relatively insensitive to the mass point, the McC estimator reacts quite strongly to the mass point as evidenced by the upwards tick just left of the cutoff. Panels C and D show the estimated densities without the mass point. In these cases, the estimates from the CJM and McC density tests are similar and overlap almost completely. These results suggests that the McC estimator reacts more strongly to such mass points.

The further simulation results in Table 2 confirm this conjecture. Using the default bins, the McC estimator starts to systematically reject the null hypothesis when a mass point of couples with equal incomes is present. Adding just 20 observations to each DGP leads to a substantial increase in the rejection rate, and the rejection rate increases rapidly with the size of the mass point. Including the mass points affects CJM's rejection rate much less, though it also exceed the nominal level once we add 100 or 200 additional observations at the cutoff value.

5. Conclusions

We show that BKP's finding of a sharp drop in the relative income distribution within married couples at the point where wives start to earn more than their husbands is unstable across different specifications. Building on a successful replication of BKP's result, our extensions for the US show that the discontinuity in relative income shares remains robust when using smaller bin sizes, but decreases substantially and becomes statistically insignificant when applying the nonparametric CJM estimator.

We then extend the analysis to administrative earnings data from Germany, where we distin-

[10] Due to the rounding rules for external researchers, we do not use the SIPP data for this exercise.

guish between married couples living in West and East Germany. For West Germany, we find a sharp drop in the relative income distribution using the BKP approach with 5% bins. The drop decreases substantially with smaller bins and becomes statistically insignificant using the CJM estimator. For East Germany, we find evidence for a drop in the relative income distribution.

Our simulation exercise shows that large, i.e., 5%, bins threaten the validity of McC's estimator as we find massive overrejection in some settings. Any conclusions should therefore be based on McC with smaller bins or on CJM. These two estimators, however, yield different findings in the US as well as in West Germany.

The divergence of these two estimators is puzzling, as both estimators provide consistent estimates of the size of the discontinuity relying on the same key assumption that several derivatives of the density function exist. The mass point of couples with identical incomes violates the smoothness assumption of both estimators. When excluding these couples, we find similar point estimates for both estimators, and the estimates are consistently lower than previously. For the US, the drop is on the margin of statistical significance when using the McC estimator, but insignificant when using the CJM estimator. For West Germany, it is insignificant throughout, and for East Germany the drop remains significant when using the McC estimator but not when using the CJM estimator.

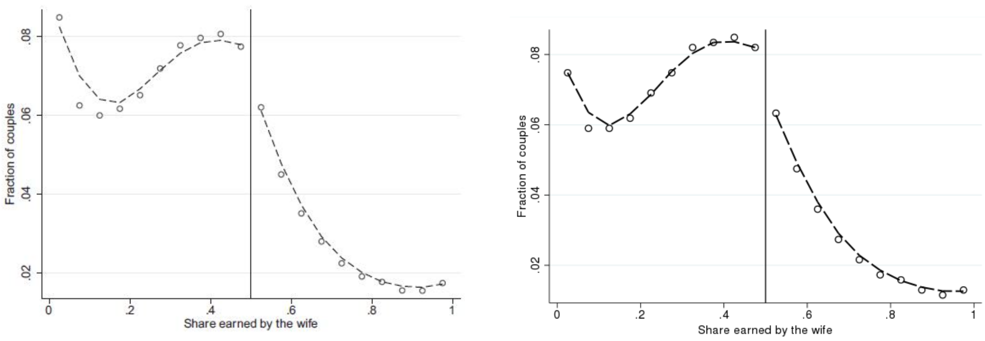
The convergence of the two estimators once we drop couples with identical incomes suggests that the McC estimator reacts more strongly to mass points than the CJM estimator. A closer graphical inspection and a simulation confirm this conjecture. How to interpret and treat the mass point of couples with identical incomes in the analysis remains a substantive question beyond the scope of our paper. The interpretation of such couples also affects the choice of the estimator – in particular if a close inspection of the underlying data is not feasible, e.g., for data privacy reasons. If one considers their identical earnings as a manifestation of gender norms, their existence should increase the rejection rate. This would favour McC. If one attributes their equal earnings mainly to other causes, e.g., labour market institutions, their impact on the rejection rate should be limited. This would point towards CJM. Further, a direct implication for future research is to inspect the data for mass points near the supposed discontinuity using finely grained data.

Taken together, our results are informative both about how couples behave in different contexts – the US, West Germany and East Germany – and how the McC and CJM estimators behave when their underlying assumptions are either met or violated. If two similar estimators produce substantially different results, researchers should obviously be concerned and try to understand the source of the discrepancy. Our paper carries the broader lesson that several alternative specifications and ideally more than one estimation procedure should be presented in empirical work involving density estimation. While pre-binning the data provides a useful starting point, we advocate for an approach that combines graphical and statistical analyses from differently pre-binned data complemented with other procedures. More broadly speaking, many empirical applications would benefit from testing the robustness of their findings with different methods instead of reporting numerous minor modifications within the same method.

Figures and Tables

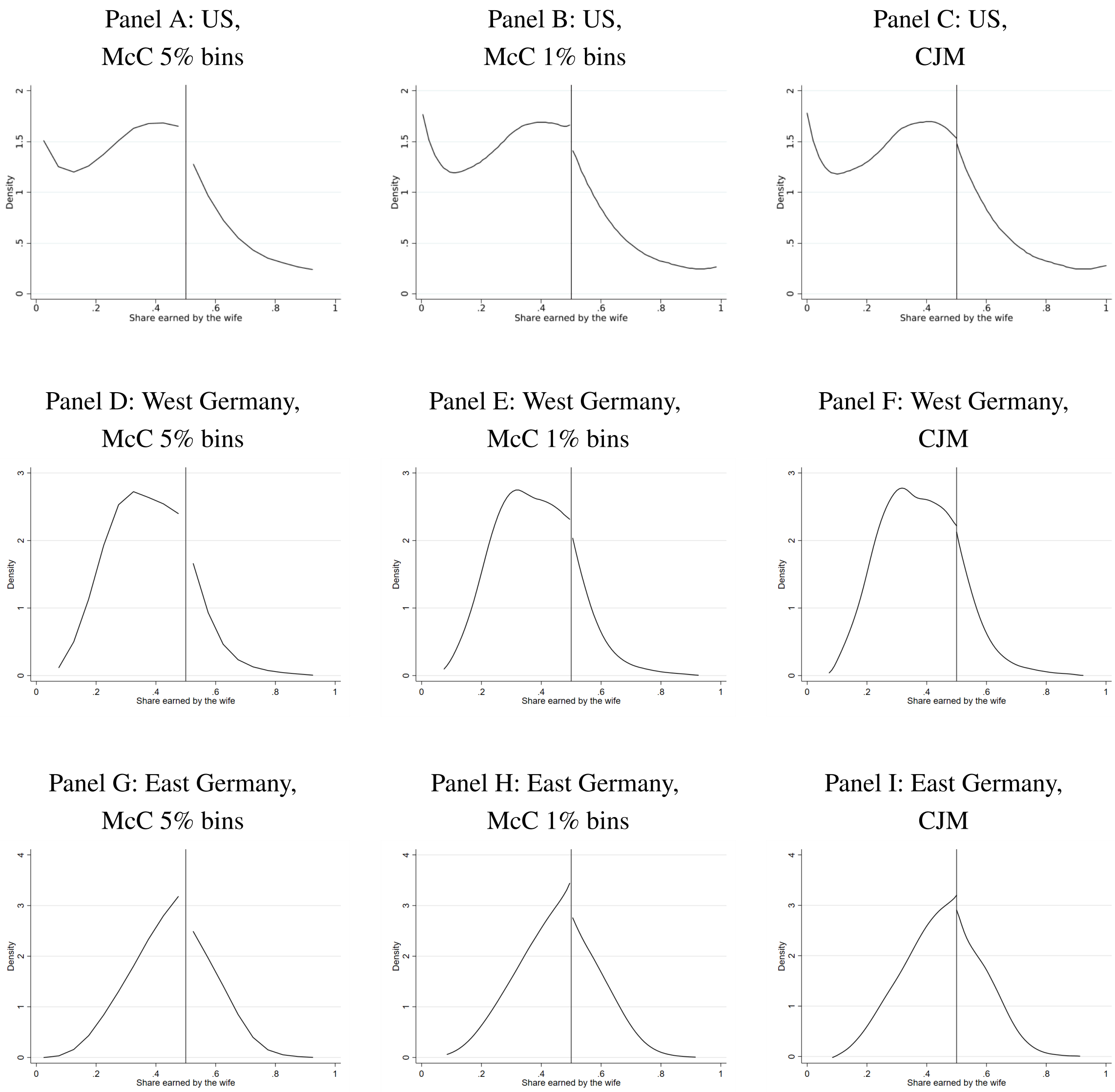
Figure 1: Distribution of relative income within couples, US data

Panel A: Original BKP graph, 5% bins | Panel B: Replication, 5% bins



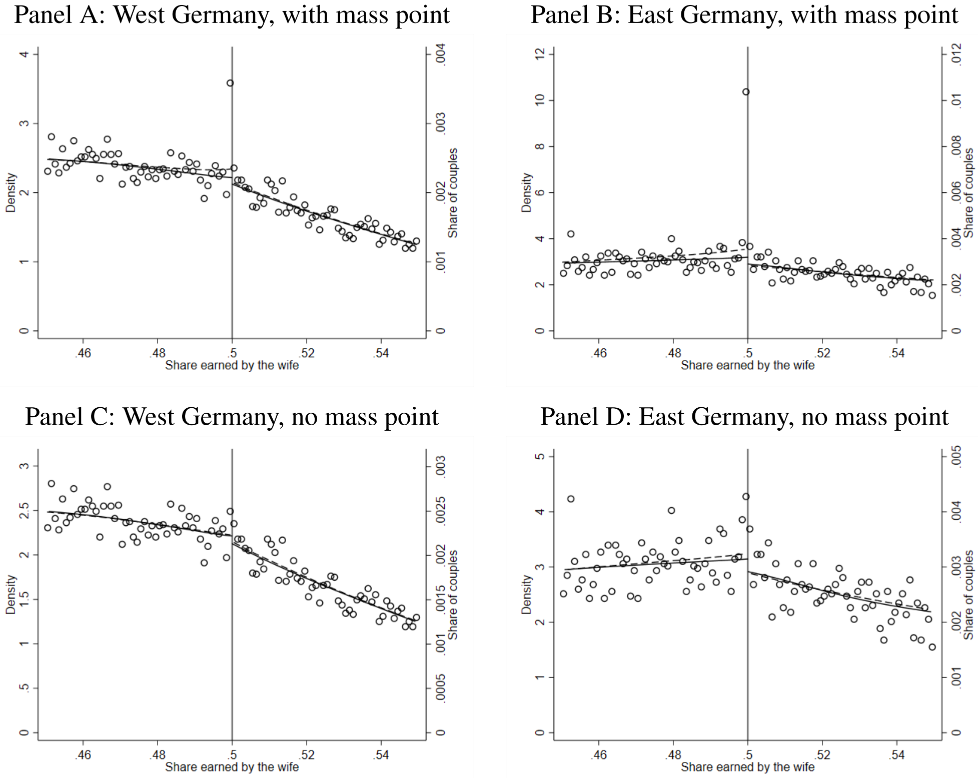
Notes: The data are from the 1990 to 2004 SIPP/SSA/IRS Gold Standard Files. For both graphs, each dot represents the fraction of couples in 5% relative income bins. The vertical line indicates the point at which the wives start to earn more than their husbands. The dashed line is the locally weighted scatter plot applied to each side of the cutoff (0.500001). The disclosure approval number for the US estimates in Panel B is CBDRB-FY20-CED001-B0006.

Figure 2: Densities of relative incomes within couples, different specifications



Notes: The data from Panels A to C are from the 1990 to 2004 SIPP/SSA/IRS Gold Standard Files for the US. The data from Panels D to I are from the Integrated Employment Biographies (IAB Integrierte Erwerbsbiographien (IEB) V12.01) for Germany. The vertical line indicates the point at which the wives start to earn more than their husband. The solid line is the estimated density at each side of the cutoff (0.500001) using the McCrary (2008, McC) estimator or the Cattaneo et al. (2020, CJM) estimator. The disclosure approval number for the US estimates is CBDRB-FY21-CED001-B0001.

Figure 3: Distribution of relative income shares in couples close to the cutoff



Notes: The data are from the Integrated Employment Biographies (IEB) for Germany and focuses on couples close to the cutoff (0.500001). Each dot represents the fraction of couples in 0.1% bins (right y-axis). The solid line displays the densities of the Cattaneo et al. (2020, CJM) estimator and the dashed line the densities of the McCrary (2008, McC) estimator using default bins (left y-axis).

Table 1: Discontinuity estimates

	McC estimator with			CJM estimator	
	5% bins	1% bins	default bins		
Discontinuity in	log density	log density	log density	density	log density
Panel A: Original results, BKP	-.123	n.r.	n.r.	n.r.	n.r.
(N=73,654)	p<0.01				
Panel B: Replication for the US					
All couples (N≈69,500)	-.119	-.140	-.153	-.101	-.067
	(4.135)	(5.014)	(4.716)	(1.181)	
excl. mass point at 0.5 (N≈69,000)	-.080	-.066	-.064	-.058	-.039
	(2.772)	(2.316)	(1.927)	(.657)	
Panel C: West Germany					
All couples (N=86,159)	-.369	-.075	-.087	-.109	-.049
	(12.850)	(2.570)	(2.815)	(.950)	
excl. mass point at 0.5 (N=86,065)	-.360	-.037	-.033	-.092	-.042
	(12.490)	(1.253)	(1.058)	(.786)	
Panel D: East Germany					
All couples (N=23,994)	-.199	-.214	-.222	-.487	-.157
	(5.332)	(6.337)	(6.469)	(2.230)	
excl. mass point at 0.5 (N=23,847)	-.143	-.118	-.114	-.311	-.103
	(3.743)	(3.217)	(3.019)	(1.366)	

Notes: This table summarises the sample sizes and estimates for different estimators and specifications. The cutoff is equal to 0.500001. Absolute t-statistics reported in parentheses. Panel A reports the original results from BKP, where n.r. indicates that BKP did not report these results in their original publication. Panel B reports the results of our replication for the US using the SIPP/SSA/IRS Gold Standard Files. Due to the disclosure guidelines of the US Census Bureau, we have to report a rounded sample size. Panels C and D re-do the analyses for Germany using the IEB data. The McC estimator uses the log density, the CJM estimator the absolute density. To facilitate comparing the size of the discontinuity estimates, we additionally report discontinuity estimates for the CJM estimator in logs. We compute the difference in the log density from the CJM estimator as $\log(\hat{f}_{right}) - \log(\hat{f}_{left})$. *The disclosure approval number for the US estimates is CBDRB-FY21-CED001-B0001.*

Table 2: Simulation for McC and CJM

	US (N=69,000)				West Germany (N=70,000)				East Germany (N=24,000)			
	McC			CJM	McC			CJM	McC			CJM
Mass point	bin size			default	bin size			default	bin size			default
at 0.5	5%	1%	default		5%	1%	default		5%	1%	default	
no extra obs.	0.951	0.048	0.050	0.050	0.936	0.089	0.059	0.053	0.058	0.045	0.042	0.055
20 obs.	0.958	0.077	0.080	0.050	0.951	0.147	0.100	0.052	0.067	0.060	0.066	0.055
50 obs.	0.966	0.192	0.237	0.053	0.969	0.267	0.247	0.052	0.102	0.136	0.174	0.056
100 obs.	0.977	0.545	0.692	0.072	0.988	0.522	0.628	0.054	0.201	0.379	0.534	0.076
200 obs.	0.994	0.984	0.998	0.287	0.998	0.935	0.990	0.087	0.554	0.920	0.974	0.248

Notes: This table reports the rejection rates (in %) of the null hypothesis of no effect based on 2,000 replications each. The cutoff is equal to 0.500001. For further details, see Online Appendix A.3.

Reference

Becker, G. S. (1973). A Theory of Marriage: Part I. *Journal of Political Economy 81*(4), 813–846.

Bertrand, M., E. Kamenica, and J. Pan (2015). Gender Identity and Relative Income within Households. *The Quarterly Journal of Economics 130*(2), 571–614.

Binder, A. J. and D. Lam (forthcoming). Is There a Male Breadwinner Norm? The Hazards of Inferring Preferences from Marriage Market Outcomes. *Journal of Human Resources*.

Cattaneo, M. D., M. Jansson, and X. Ma (2018). Manipulation Testing Based on Density Discontinuity. *The Stata Journal 1*, 1–18.

Cattaneo, M. D., M. Jansson, and X. Ma (2020). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Codazzi, K., V. Pero, and A. Albuquerque Sant'Anna (2018). Social Norms and Female Labor Participation in Brazil. *Review of Development Economics 22*(4), 1513–1535.

Hederos Eriksson, K. and A. Stenberg (2015). Gender Identity and Relative Income within Households: Evidence from Sweden. IZA Discussion Paper 9533, IZA Bonn.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics 142*(2), 698–714.

Murray-Close, M. and M. L. Heggeness (2018). Manning Up and Womaning Down: How Husbands and Wives Report Their Earnings when She Earns More. Working Paper 2018-20, Social, Economic, and Housing Statistics Devision, US Census Bureau.

NYT (2018). The hidden taxes on women. 02-03-2018.

Roth, A. and M. Slotwinski (2018). Gender Norms and Income Misreporting within Households. Working Paper 7298, CESifo.

Schnabel, C. (2016). United, yet Apart? A Note on Persistent Labour Market Differences between Western and Eastern Germany. *Jahrbücher für Nationalökonomie und Statistik 236*(2), 157–179.

Sprengholz, M., A. Wieber, and E. Holst (2019). Gender Identity and Wives' Labor Market Outcomes in West and East Germany between 1984 and 2016. DIW Working Paper 1030, DIW Berlin.

The Economist (2012). When Women Dare to Outearn Men. 18-12-2012.

UNDP (2020). *Tackling Social Norms. A Game Changer for Gender Inequalities*. New York: United Nations Development Programme.

Zinovyeva, N. and M. Tverdostup (2018). Gender Identity, Co-Working Spouses and Relative Income within Households. IZA Discussion Paper 11757.

Appendix A. Online Appendix

Appendix A.1. Adjustments in Stata ado-file DCdensity

Our estimations of McCrary (2008)'s density test build upon the user-written Stata ado-file *DCdensity*.[1] The ado-file is tailored towards a running variable with a distribution function that thins out towards extreme values. The share of the wives' income in the US data, in contrast, does not necessarily thin out towards the boundaries of 0 and 1. To improve the density estimation close to the boundary points 0 and 1, we made three minor adjustments to the ado-file. These adjustments affect the density estimates for income shares close to 0 and 1 and thus the graphical representation of the density, but they are practically irrelevant for the discontinuity estimates.

First, the original ado-file smooths the distribution at the boundaries of the observed range of the running variable. Specifically, it adds bins that contain no observations below the lowest bin and above the highest bin. In our application, these bins can lie (partly) outside of the unit interval and would lower the density estimates close to 0 and 1. We therefore do not add such bins.

Second, *DCdensity* bases the construction of bins on the range of observed values of the running variable. As this range differs across the implicates of the SIPP data we work with, we base the construction of bins on the unit interval, which is the range of possible values in our analysis.

Third, the original choice of the number of bins and their midpoints ensures that all observed values fall within a bin. It does so by allowing for bins that lie (partly) outside of the observed range of values or in our case the unit interval. In our setting, such bins lower the density estimates close to 0 and 1 as there cannot be any couples with income shares below 0 or above 1. To avoid this issue, we only use bins that fully lie in the unit interval. This comes at the cost of potentially omitting couples in which the wife earns an income share close to 0 or to 1 (more precisely, below the lower bound of the lowest bin or above the upper bound of the highest bin).

The augmented ado-file is part of the supplementary material to the article.

Appendix A.2. Details for German data

We use administrative data from the social security system, the Integrated Employment Biographies (IAB Integrierte Erwerbsbiographien (IEB) V12.01; see Jacobebbinghaus and Seth, 2007, for a description). The IEB contain information on individual employment starting from 1975 for all employees who are subject to social security contributions. As the employer-reported gross earnings are used to calculate contributions to and benefits from the social security system, information on incomes and socio-demographic characteristics (e.g., age and gender) are

[1] The ado-file was written by Brian Kovak and we use the version as of 2009, which is for instance available at `https://eml.berkeley.edu/~jmccrary/DCdensity/`. Last accessed March 11th 2021.

extremely reliable. Mirroring BKP's sampling restrictions, we include only married couples in which both partners earn a positive income and are 18 to 65 years old in our sample.

As the administrative data contain no direct information on marital status or partners, we use the procedure developed by Goldschmidt et al. (2017) to identify married couples. This method identifies individuals as couples if they are the only two persons living at the same address who share the same surname and if their age difference is at most 15 years. Marrying couples can either adopt the groom's or the bride's surname, an overlapping double surname, or groom and bride can keep their names. As the couple identification is based on shared surnames, it cannot identify couples in which both partners keep their names. However, according to Gesellschaft für deutsche Sprache (2018), only roughly 12% of German couples do so and this limitation is hence no major issue in identifying married couples. The method identifies couples in which both partners are registered with Germany's Federal Employment Agency, as employed subject to social security or in marginal employment, unemployed, registered as job-seekers or in a labour market program, as of 30 June 2008. We therefore examine relative income within couples as of this date. See Goldschmidt et al. (2017) for detailed information on the validity of the derived couples indicator. We use a 10% sample of all identified couples.

Due to the institutional setting and the data, the analysis for Germany differs in three aspects from the analysis for the US. First, the German data comprise all individuals who are subject to social security contribution, i.e., they do not cover self-employed individuals and civil servants. Second, the earnings data is top-coded at the social security contributions ceiling. This prevents us from computing the exact income shares of both partners if at least one partner's income lies above this threshold. To deal with this issue, we exclude couples in which at least one partner earns more than the social security contributions ceiling. Third, the German tax and social security system requires lower contributions from workers who earn less than 400 Euro per month (in 2008), which creates an excess mass of jobs with monthly earnings at and below this threshold.[2] To avoid that this threshold affects our results, say because the bunching of earnings affects the overall distribution of income shares or both partners earn exactly 400 Euro, we exclude couples in which one or both partners are marginally employed.

Appendix A.3. Description of simulation

For our simulation exercise, we use three different DGPs. These processes all rely on the beta distribution. The beta distribution appears well-suited, because it is defined over the unit interval (as is the wife's income share) and can flexibly be modelled using two shape parameters. The combination of these shape parameters determines the center and the spread of the density function. To fit the simulated data to the empirical distributions in the US, West Germany and

[2] Up to this threshold, earnings are subject to total contribution rate of 30 % for social security and income taxation paid by the employer. If the earnings exceed this threshold, employers and employees pay social security contributions on the order of 20% each and the employee is additionally subject to standard income taxation. More information on marginal employment is available in Collischon et al. (2020).

East Germany, we combine several beta distributions with different parameters. In addition, we use a sample sizes that are similar to our actual samples. The details of the DGPs are as follows.

Data generating process "US" with $N_{sim} = 69,000$

$$x \sim \begin{cases} Beta(4,7) & \text{with prob. } 80\% \\ Beta(1,15) & \text{with prob. } 15\% \\ Beta(8,1) & \text{with prob. } 5\% \end{cases}$$

Data generating process "West Germany" with $N_{sim} = 70,000$

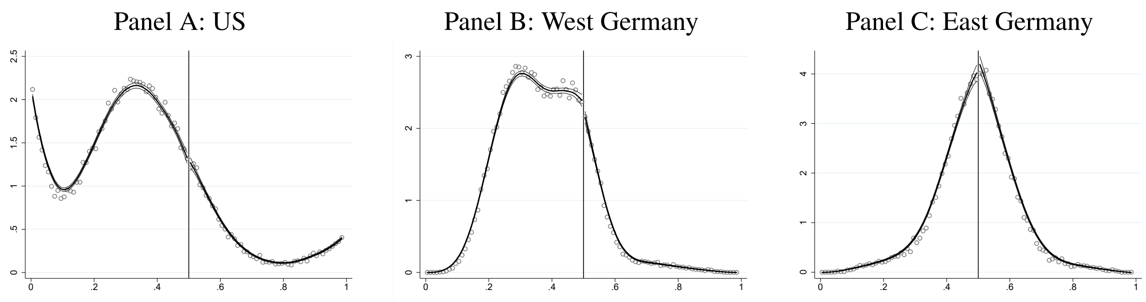
$$x \sim \begin{cases} Beta(7,15) & \text{with prob. } 70\% \\ Beta(34,36) & \text{with prob. } 25\% \\ Beta(7,4) & \text{with prob. } 5\% \end{cases}$$

Data generating process "East Germany" with $N_{sim} = 24,000$

$$x \sim \begin{cases} Beta(15,15) & \text{with prob. } 75\% \\ Beta(4,8) & \text{with prob. } 10\% \\ Beta(25,25) & \text{with prob. } 10\% \\ Beta(10,4) & \text{with prob. } 5\% \end{cases}$$

Using each DGP, we simulate 2,000 data sets. To explore the influence of couples with identical incomes, we add 20/50/100/200 observations with $x = 0.5$ to each of these data sets. Figure A.1 below depicts one draw for each DGP alongside the McCrary (2008) density estimates for these draws before adding couples with identical incomes.

Figure A.1: Simulated densities of relative incomes within couples



Notes: Simulated densities as described in the text and McCrary (2008) graphical density estimations.

References for Online Appendix

Collischon, M., K. Cygan-Rehm, and R. T. Riphahn (2020). Employment Effects of Payroll Tax Subsidies. *Small Business Economics* (forthcoming).

Gesellschaft für deutsche Sprache (2018). Familiennamen bei der Heirat und Vornamenprognose 2018. 19-12-2018.

Goldschmidt, D., W. Klosterhuber, and J. F. Schmieder (2017). Identifying Couples in Administrative Data. *Journal for Labour Market Research 50*(1), 29–43.

Jacobebbinghaus, P. and S. Seth (2007). The German Integrated Employment Biographies sample IEBS. *Schmollers Jahrbuch 127*(2), 335–342.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics 142*(2), 698–714.