The Effect of the Relative Wage on the Labor Supply of Married Women*

Sarah Rosenberg†

Abstract

Research by Bertrand, Kamenica, and Pan (2015) finds evidence of a negative association between the probability that a wife outearns her husband and her labor supply, suggestive of the "breadwinner norm." This paper reassesses this relationship by estimating the causal effect of the intra-household relative wage on married women's labor supply. Exploiting gender-based variation in the geographic distribution of industries, Bartik instruments for men and women's wages capture exogenous variation in the relative wage owing to labor demand shocks. A negative causal relationship would imply that efforts to reduce the gender wage gap could paradoxically increase the gender earnings gap: women would work less as their relative wages rise. However, my results imply that increases in the relative wage in fact increase married women's hours.

Keywords Labor supply, Household economics, Gender norms, Relative wage

JEL Codes D13, J22

*I would like to thank seminar participants at Stockholm University, ECARES, the Institute for Fiscal Studies, the Society of Household Economics Conference, and the IZA World Labor Conference 2018 for thoughtful comments and suggestions. I am grateful to Emir Kamenica for sharing the code necessary for replicating his and his co-authors' paper. My research is made possible by funding via a FRESH Grant from the F.N.R.S.

†Université Libre de Bruxelles, ECARES. Email: srosenbe@ulb.ac.be

1 Introduction

Most married women in opposite-sex couples earn less than their husbands on an annual basis, even when both are working. While women are more likely to be the primary earner than in the past, this designation stills falls to the husband for more than two-thirds of American couples in recent years. A natural explanation for this difference might be that women on average have lower wages than men. After all, the gender wage gap persists despite having shrunk (Blau and Kahn, 2017; Goldin, 2014).This explanation invokes the notion of comparative advantage with respect to wages: we expect the higher-wage individual to work and earn more in the market.

In this paper, I explore whether the pattern of primary earners can be in part explained by a causal impact of the breadwinner norm on women's labor supply within marriage. The breadwinner norm, first discussed in the economics literature by Bertrand et al. (2015) (hereafter BKP), is the principle that husbands should earn more than wives. Concretely, I consider whether an increase in the relative potential earnings of the wife causes women to decrease their hours at the intensive margin. Such an effect would be consistent with one interpretation of a key finding in BKP: a negative association between a measure of the likelihood of a wife outearning her husband and her labor supply. While acknowledging that their analysis is not causal, they interpret this evidence as suggestive that some women may leave the labor force or decrease their labor supply when the probability of outearning their husbands is high.

The contribution of this paper is to estimate the causal effect of changes in relative potential earnings on married women's labor supply. If the breadwinner norm affects married women's labor supply behavior, efforts by policymakers to promote women's wages in order to reduce gender earnings gaps may backfire. Higher wages increase the probability that married women would outearn their husbands. If the norm has a direct causal effect on labor supply, women may respond to higher relative potential earnings by decreasing their hours, and thereby decreasing their earnings.

I use relative hourly wages as a measure of relative potential earnings. Hourly wages are the basic measure of potential earnings used in many labor supply analyses. Hourly wages have the advantage of being observed for all individuals who are working. At the same time, measurement error in hourly wages presents potential issues.

To begin my analysis, I show descriptively some key facts that are suggestive of the breadwinner norm. Using data from the U.S. Census and from the Panel Study of Income Dynamics, I observe higher-wage wives (compared to husbands) for a larger fraction of couples than higher-earning wives. This difference suggests some higher-wage wives work fewer hours than their husbands. Illustrated differently, when I relate the intra-household relative hourly wage to the annual hours of wives and husbands for couples that are both working, wives' average hours appear to rise until the point at which the wife and husband's hourly wages are equal, after which they sharply decline–an inverse U-shaped relationship strikingly reminiscent of the breadwinner norm.

There are two key weaknesses in interpreting the descriptive relationships as evidence of the breadwinner norm. First, measurement error for wages is unaccounted for. Second, even if measurement error is not an issue, it may not be representative of the causal relationship between the relative wage and hours.

To account for measurement error, I instrument for current-period relative wages using lagged relative wages with panel data from the PSID. I find no evidence of an inverse U-shaped relationship. Instead, the relationship appears better approximated as positive and linear.

Since the intra-household relative wage is endogenous, to actually estimate the causal effect of the relative wage I employ a Bartik instrument for men and women's wages respectively. These instruments capture variation in the relative wage that owes to gender-specific differences in labor demand at the national level, which in turn differentially affect local labor markets based on their initial exposure to higher- or lower-growth industries.

Using this instrumental variables strategy, I find no evidence that couples respond to an increase in the relative wage owing to labor demand shocks by decreasing wives' labor supply. When the relationship between married women's hours and the relative wage is estimated linearly, the estimated causal effect is positive and significant. When I allow for the possibility of nonlinearity by including a quadratic term, the marginal effect of the relative wage becomes either positive or highly imprecise across the distribution of the relative wage.

As a result, I do not find evidence that "breadwinner norm" affects labor supply decisions *within* marriage. Rather, the causal effect of the relative wage on hours appears to be positive. Given that BKP also find evidence that (potential) wives earning more may lower marriage rates and increase the probability of divorce, the breadwinner norm may instead have some impact on selection into and out of marriage, which would also be consistent with the results of Folk and Rickne (2019).

I focus on the intensive margin for two reasons. First, it circumvents the difficulty of predicting relative wages for couples with a non-working member. Secondly, if the breadwinner norm were to have a direct impact on the extensive margin, I would certainly expect to see an impact on the intensive margin; reducing hours is a much less costly way to avoid violating the norm than fully dropping out of the labor force.

These findings contribute to the literature on societal gender norms and labor supply. While at the macro level, there is clearly an association between traditional gender norms (typically measured by national surveys) and lower female labor force participation (Fernández, 2013; Fortin, 2005, 2015)), at the individual or household level it is difficult to distinguish between behavior that appears consistent with a particular gender norm (as in the reduced-form analysis of BKP) and other possible explanations.

In this spirit, a small but growing literature explores which mechanisms may drive effects that seem consistent with gender norms and which can be ruled out, focusing on the individual and couple level. Most relevant to this paper, two papers find in experimental settings that couples choose specialization primarily based on efficiency (productivity, or relative costs of contributing to a public good) rather than gender identity concerns (Görges, 2018; Cochard, Couprie, and Hopfensitz, 2018). This paper complements the experimental approach in these articles by testing for the predicted causal impact of the breadwinner norm using observational data.

More broadly, given BKP's wide array of intriguing results motivating their claims about the breadwinner norm, several papers have reconsidered both whether their results can be interpreted as evidence of the breadwinner norm and whether their findings hold in other contexts (Binder and Lam, 2018; Zinovyeva and Tverdostup, 2018; Hederos Eriksson and Stenberg, 2015; Foster and Stratton, 2018b). This paper adds to this area as the first to reexamine BKP's results about the breadwinner norm with respect to labor supply.

Finally, this paper is also contributes to research beyond the breadwinner norm in considering how shocks to gender roles in marriage affect household behavior and marital outcomes. For example, Foster and Stratton (2018a) show that job promotions or losses affect the household divison of labor; Avdic and Karimi (2018) show that increases in the use of paternity leave due to changes in incentives increase marital instability; and Folk and Rickne (2019) show that women's promotions to top jobs increase marital instability, particularly if it entails a substantial reversal of relative earnings. Understanding the effects of shocks within marriage may be important to anticipating the effects of policies intended to influence gender roles. It can also give insight to the extent to which selection into marriage as opposed to changes after marriage matter.

The paper proceeds as follows. Sec. 2 describes the data used and illustrates the apparent inverse U-shaped relationship between wives' hours and the intrahousehold relative wage and Sec. 3 describes the empirical framework. Results are discussed in two parts, with Sec. 4 addressing the role of measurement error using the PSID and Sec. 5 estimating the causal effect of the relative wage using Bartik instruments. Sec. 6 discusses the findings and Sec. 7 concludes.

2 Relative Wage and Wives' Hours: An Inverse U-shaped Relationship?

In this section, I present descriptive evidence on the distribution of the intra-household relative wage and its relationship with married individuals' hours to motivate the subsequent analysis.

2.1 Data

I use two sources of data in this paper: the U.S. Census (including American Community Survey waves for more recent years), accessed via IPUMS (Ruggles, Flood, Goeken, Grover, Meyer, Pacas, and Sobek, 2018) and the Panel Study of Income Dynamics (??, 2018).

For the Census, I use the 5% Public Use Microdata for the years 1980, 1990, and 2000. I also use the American Community Survey waves 2005-2007[1], which I pool and refer to as 2006 in the remainder of the paper. Hourly wages in the Census are calculated by dividing annual wage income by weeks worked multiplied by usual weekly hours.

For the PSID, I use responses from married couples from all years available, from 1969-2013. In this dataset, only annual hours are reported, so hourly wages are calculated by dividing annual wage income by annual hours.

For my sample, I include married couples between the ages of 18 and 65 where neither member is retired or enrolled in school. Much of the analysis focuses on couples where both work positive hours in the market. Tables 1, 2 provide descriptive statistics on the characteristics of married couples in both datasets.

[1] I use these waves because 2007 is the last year that individuals were asked an exact number of weeks worked rather than intervalled responses, which allows me to calculate hourly wages.

2.2 Descriptive Evidence

In BKP, the focus throughout is on measures of relative income. Arguably, the intuition is that the breadwinner norm operates with reference to annual income, not wages. Perhaps because of this, their measure for "the probability that a wife earns more" than her husband involves assigning all women potential earnings from the distribution of earnings of demographically similar women who are working, whether full-time or part-time. In other words, each woman's potential earnings is defined by the joint distribution of hours and earnings of similar women. It is not straightforward to consider in what direction this inclusion of hours might bias estimation of the relationship between labor supply and the measure of potential relative earnings. For women with many observably similar peers working full-time, it may bias upwards their potential earnings, while the opposite would be true for demographic groups where many work part-time.[2]

Given this ambiguity, I begin by returning to the most basic measure of potential wages used in labor economics: the hourly wage, given by dividing labor income by hours worked. I limit the sample to couples where both members are working positive hours so that I can directly calculate their wages. Figs. 1 and 2 compare the distribution of intra-household relative (labor) income to the distribution of relative hourly wages for the Census/ACS and PSID over time, excluding self-employed individuals.[3] The relative wage and income measures are calculated analogously as $\pi = \frac{\omega_w}{\omega_w + \omega_m}$ where w stands for women and m for men, and ω_i is earnings or wages respectively.

Beginning with the distribution of relative income, it's clear that across years

[2] In Appendix A I replicate the results from BKP and consider modifications of their key variable, intended to capture the probability that a wife could earn more than her husband. I find that both the sign and magnitude of the main coefficient are highly sensitive. The modifications I test are based on the insight that for women who in actuality earn more than their husbands, BKP's probability measure is arguably biased downwards, assigning them on average a 36% probability of earning more. I precisely define and discuss the original measure used in BKP, my adapted measures, and show results for the replication and alternative specifications. The sensitivity of the results to ostensible bias in the key measure provides additional motivation for examining the impact of the relative wage.

[3] BKP include self-employment and non-wage-income in their depictions of relative income, but I exclude them as both aspects present issues when included in the estimate of hourly wages.

and both datasets, most of the mass is to the left of 0.5: Unsurprisingly, most women earn less than half of the household's income. Over the years, there is a shift towards greater concentration of couples in the 0.3 to 0.5 range, and an increase in the fraction of couples with a higher-earning wife, though they still remain in the minority. Another notable feature that becomes more striking in the most recent years (2000 and 2006 in the Census/ACS; 2000s and 2010s for PSID) is the spike that appears quite precisely near 0.5. This spike remains after correcting for rounding issues in the ACS/Census (see Appendix B for details) and removing top-coded incomes. One reason it might appear appear more prominent in my graphs is that, for the purposes of my analysis, I exclude the self-employed, since it is difficult to gauge whether their reported hours are accurate. Recently, both Zinovyeva and Tverdostup (2018) and Hederos Eriksson and Stenberg (2015) have suggested the mass at 0.5 is at least in part due to couples where both work in the same sector, which may in fact indicate something like joint ownership of a family business.

While this spike is not the main point of interest for me, it should be noted that it figures importantly in BKP and the subsequent discussion for the breadwinner norm: BKP argued that this discontinuity is evidence that couples are bunching below this threshold to avoid violating the norm.[4]

More relevant for my purposes is the broader differences between the distribution of relative income and wages. Whereas the relative income is sharply skewed to the left, the distribution of relative wages has its mode just to the left of 0.5 and appears to be quite normally distributed. A greater part of the sample is concentrated between 0.3 and 0.7 for the PSID compared to the Census/ACS, perhaps because these pooled observations for the PSID in actuality contain repeated measures for the same individuals, but it nevertheless also appears normal. Considering the right half of the distributions for the ACS/Census, there is a "missing mass" between

[4]This claim has generated additional follow-ups beyond the two just mentioned. One article has also argued that even if there is a discontinuity, it is consistent with a range of other preferences aside from aversion to the wife earning more (Binder and Lam, 2018). Finally, a Census analysis comparing Census-reported earnings to administrative earnings data and found that in couples where the wife actually earned more, her income was systematically underreported by 1.5% and the husbands overreported by 2.9% (Murray-Close and Heggeness, 2018).

the two distributions: a fraction of women have a higher hourly wage than their husbands and yet earn less.

This apparent gap between wage and earnings suggests that some women may be working less than their husbands despite having a higher wage. To examine the relationship with hours, Fig. 3 and Fig. 4 graph the average hours worked by men and women respectively by quantiles of the relatively hourly wage distribution and by decade. The pattern for women is striking and consistent across the two datasets and across time: the average hours rise across the distribution until close to 0.5, where husband and wife have equal wages, and after this point the average hours decline. Although average hours for women rise over time, it does so across the distribution, so that the inverse U-shaped relationship persists.

For men, though there is also somewhat surprisingly a U-shaped (or perhaps better described as M-shaped) relationship, it's important to note that the range of hours is much smaller than observed for women: only a difference of about 300 hours between the lowest levels around 2000 hours per year to 2300. In other words, though there is some variation,men seem to work full-time across the distribution of relative wages. In comparison, women's hours exhibit ranges of between 500 and 800 hours depending on the year and dataset.

On its face, these graphs go in line with the story of the breadwinner norm. In fact, the relationship appears to be stronger than that estimated in BKP, which was quite small in economic terms. That the change in slope occurs close to 0.5 is particularly striking with the breadwinner norm in mind.

Interpreting these patterns as evidence of the breadwinner norm depends crucially on the relative hourly wage as a consistent measure of a couple's relative earnings potential. If this measure is biased, the graphs may misleadingly suggest support for the breadwinner norm. Beyond that, these descriptive relationships may not be indicative of the causal effect of the relative wage. In the remainder of the paper, I address the key issues of measurement error and endogeneity of the relative wage.

3 Empirical Framework

The key goal of this paper is to test whether the breadwinner norm might indeed cause women to work less when they are likely to earn more than their husbands, while accounting for factors that might spuriously lead to inferring the impact of the norm.

The relative hourly wage as the main measure of relative potential earnings maps straightforwardly to thinking about the breadwinner norm: when the wife has a lower wage than her husband, her probability of earning more than her husband is negligible. If earnings roughly linearly with hours, where she has a lower wage she cannot earn more even if she were to work full-time. When her wage exceeds his, the probability that she could earn more is positive and increasing in her wage.[5] Thus, I am interested in estimating equations of the following form:

$$h_{wt} = \Pi_{kt}\alpha + X_{kt}\beta + [\mu_k] + \gamma_t + u_{kt} \tag{1}$$

where the household is denoted by k and Π_{kt} is a vector of relative wage variables and $\pi_{kt} = \frac{\omega_{wt}}{\omega_{wt}+\omega_{mt}}$ is the intra-household relative wage. Given the inverse U-shaped relationship described in Sec. 2, I consider both linear and quadratic specifications of π_{kt}. X_{kt} is a vector of observable variables that may be correlated with the relative wage, μ_k are couple fixed effects (where possible), γ_t time fixed effects, and u_{kt} represents the unobserved component specific to the couple in a given time period.

The interpretation of α depends on the source of the variation in Π_{kt}. Estimating any relationship between hourly wages and labor supply brings up two well-known issues: measurement error and the endogeneity of wages. The goals of the empirical analysis are to consider the relationship between Π_{kt} and h_{wt} while accounting for measurement error, and to estimate the causal effect of an increase in the relative

[5] There is evidence for a wage penalty for part-time work (Bardasi and Gornick, 2008; Hirsch, 2005; McGinnity and McManus, 2007; Manning and Petrongolo, 2008), thus the full-time wage an individual could attain might in fact be greater than the linear extrapolation of a part-time wage. However, this would merely have the effect of shifting the point at which the probability of the wife earning more is positive to a slightly lower relative wage.

wage. If the breadwinner norm has a causal effect on labor supply, I would expect that the marginal effect of the relative wage is negative for high levels of the relative wage (e.g. when the wife has a positive probability of earning more than the husband if she were to work full-time.)

3.1 Accounting for measurement error

The observed relative wage π_{kt} is a biased measure of the true relative wage owing to the presence of measurement error in the observed wages of each household member. Typically it's assumed that the source of measurement error in hourly wages owes to error in hours, rather than earnings. As a result, the measurement error enters both the independent and dependent variables. In a standard estimation of wage-hours elasticities, Borjas (1980) shows that this errors-in-variables problem leads to a downward bias in the coefficient on wages, known as division bias. This problem of measurement error is the one generally discussed in labor supply analyses.

It's not obvious that measurement error plays the same role in this analysis, where there may be measurement error in both wives' and husbands' wages. However, as Griliches and Hausman (1986) point out in their seminal article on errors-in-variables in panel data, if one is willing to assume that the error terms are serially uncorrelated, a straightforward solution is to instrument for the variable in question using lags.

In this case, using panel data from the PSID, I can instrument for the true relative wage using lagged wages and thereby test whether the inverse U-shaped relationship described in Sec. 2 still holds when excluding the effect of measurement error.[6]

[6] Standard methods used to address measurement error in labor supply analysis often employ group-level methods such as using group means as instruments or conducting the analysis at a group level, typically defined by age and education (Blau and Kahn, 2007; ?). These methods seem less well-suited to my setting than using panel data. Using group means as instruments is problematic as it requires imposing potentially implausible exclusion restrictions, as discussed in Blau and Kahn (2007). Conducting the analysis at a group level would present other issues. If I were to take group-level means of the intra-household relative wage, I would need to make assumptions about the distribution of measurement error for men and women's wages respectively to claim that the group-level means are purged of error. If I first take group-level means for women and men's wages respectively and then construct the relative wage, then the resulting "relative wage"

Suppose that $\widetilde{\omega}_{nt} = \omega_{nt} + \phi_{nt}$, where $\widetilde{\omega}_{nt}$ is the observed wage of individual n at time t, ω_{nt} is her true wage, and ϕ_{nt} represents measurement error. Given the assumptions that ϕ_{nt} is serially uncorrelated over time within individuals and that ϕ_{wt} and ϕ_{mt} are uncorrelated, I can use $\pi_{kt-1} = \frac{\widetilde{\omega}_{wt-1}}{\widetilde{\omega}_{wt-1}+\widetilde{\omega}_{mt-1}}$ to instrument for the true relative wage. The implied exclusion restriction is that the previous period's relative wage only affects the current period hours worked via its impact on the current period relative wage. At the same time, I also include total wages as an endogenous variable, to control for the fact that a past-period increase in wealth via total wages may directly affect current period hours via an income effect.

Is serially uncorrelated measurement error a reasonable assumption? It is often assumed to be serially uncorrelated, such as in life cycle models of labor supply, but the possibility remains that it is not. There are of course other methods for errors-in-variables models with serially correlated errors, but in lieu of going further along this route, I will compare the analysis using lagged wages in panel data to using Bartik instruments in cross-sectional data, discussed in the next section. The Bartik instruments defined at an aggregate level are orthogonal to individual measurement error, but are likely weaker as instruments than individual lagged wages, so the panel and cross-sectional analyses provide complementary approaches.

Using the PSID, I thus estimate Equation 1 using two-stage least squares with the lagged measures of the relative wage and total wage as instruments. The first stage is:

$$\Pi_{kt} = X_{kt}\beta + (\Pi_{kt-1}\delta) + \mu_k + \gamma_t + \epsilon_{kt} \tag{2}$$

Here Π_{kt} is the vector of instrumented variables, which include the intra-household relative wage, its square when relevant,[7] and the household's total wage, which is given by the sum of the individual wages. I include the household's total wage to

would represent an aggregate relative wage rather than intra-household relative wage, which is not relevant to the question at hand.

[7]When including the square, the necessary assumption is serial independence rather than serial uncorrelatdness.

disentangle potential wealth effects from changes in the relative wage itself. I instrument for it since it also would be affected by measurement error in wages.

The controls X_{kt} include the age of wife and husband and their squares, the number of children, and non-wage income. Year and couple fixed effects are also included, given by μ_k and γ_t. I also estimate the inverse Mills ratio to include as an additional control for selection into work. I estimate single-year probit equations of labor force participation on the wage variables, the controls previously mentioned, as well as dummy variables describing the age of the youngest child in the household, with the latter functioning as the excluded variables.

The second stage is then:

$$y_{kt} = X_{kt}\beta + (\hat{\Pi}_{kt}\alpha) + \mu_k + \gamma_t + u_{kt} \tag{3}$$

The dependent variable y_{kt} is the annual hours of the wife in time t. By construction, the sample is restricted to those where both the wife and husband work in two consecutive periods, such that the relative wage and its lag are both observed.

To provide a baseline comparison, I will also estimate Eq. 3 by OLS, using the observed values of the relative wage and total wage rather than the fitted first-stage values.

The parameter α characterizes the association between wives' labor supply and the intra-household relative wage across the distribution of relative wages. In thinking about the interpretation of these coefficients, particularly for the 2SLS estimation, a few remarks are in order. First, it is primarily informative about whether the inverse U-shaped relationship described earlier in this paper owes to measurement error. Second, even if the exclusion restriction is valid, wages in the previous period are still endogenous, and therefore it clearly does not describe the causal effect of changes in the relative wage. In particular, given the selected sample of women who have been in the labor force at least two years consecutively, these individuals may represent those that are more progressive with respect to gender norms, which could bias the coefficient on the relative wage upwards from zero.

3.2 The Causal Effect of the Relative Wage

As previously noted, both individual wages and the intra-household relative wage are endogenous. In particular, unobserved ability and beliefs about norms may present issues if they are related to selection into marriage and into the labor force. Unobserved ability is a commonly discussed issue in the estimation of wage-hours elasticities. The concern is that higher-ability individuals may have higher wages and also prefer to work more, biasing the elasticity estimates upwards away from zero.

Selection into marriage poses an additional problem with respect to these factors. Suppose that some fraction of individuals do care about the breadwinner norm. Intuitively, they are more likely to form marriages with other individuals who also care about the breadwinner norm. Moreover, by nature of the norm, such couples would be more likely to include lower-wage wives matched with higher-wage husbands. In contrast, those who do not care about the breadwinner norm may be more likely to form marriages with relatively higher-wage wives. If wages are also positively correlated with both ability and hours, then the relationship between observed wages (absent measurement error) would be biased upwards due to these factors correlated with selection into marriage.

With respect to men's wages, it is less obvious that there might be such a systematic relationship between norms, their average wage levels, and selection into marriage: given the choice of two potential matches that differ only with respect to wages, I would expect that women always prefer the higher-wage individual, regardless of preferences about norms.

To test this relationship causally, I would ideally wish to observe variation in the relative wage that is uncorrelated with individual abilities and beliefs about norms.

To generate this type of variation, I employ the Bartik method (Bartik, 1991) to instrument for demand-side variation in wages for men and women respectively, using local industry shares and national-level wage growth by industry over time.

3.3 Identification

The Bartik method has been employed for labor analysis as well as a range of other topics. My application of Bartik instruments is similar in spirit to its use in Aizer (2010), who looks at the effect of the average relative earnings of women and men on domestic violence rates; Page, Schaller, and Simon (2019), who consider the effect of women and men's relative employment opportunities on child health; and the marriage market analysis in BKP, which relates relative earnings in homogamous marriage markets to the marriage rate. These analyses share a feature that distinguishes them from the canonical application of Bartik instruments in that they effectively use two instruments, one for women and one for men (though BPK combine these into one measure).

My application differs in a second way from other Bartik analyses (including the above-mentioned ones) in that my outcome of interest is at the household rather than at an aggregate level. When considering the potential effect of the breadwinner norm, it is the household-level relative wage that should be relevant: the breadwinner norm is violated when a wife in a given couple earns more than the husband, not if other wives have a high relative wage.

The recent growth in the use of "shift-share" instruments in empirical applications has spurred a corresponding surge in research focused on characterizing the conditions needed for causal indentification and inference. Identification in this setting is based on the principles laid out in Goldsmith-Pinkham, Sorkin, and Swift (2018). A key takeaway from their paper is that the exclusion restriction is generally best characterized with respect to the exposure variable, which is usually an industry share.

In this setting, demand for labor in different industries changes over time, yielding differential wage growth by industry. This industry-specific growth also varies by gender. These demand shocks, which I characterize at the national level, affect relative wages in local labor markets differently based on the distribution of men and women across industries at an initial point in time. Variation comes from both gendered differences in wage growth within an industry (at the national level) and gender segregation across industries (at the local level). Intuitively, an area that had a relatively higher proportion of women in an industry with faster-growing wages for women would be expected to have greater increases in the relative wage than another area with fewer women in this industry.

As previously described, a key challenge in estimating the causal effect of the relative wage is that sorting in the marriage market on gender norms might itself lead to a positive relationship between the relative wage and wives' hours, once measurement error is accounted for. If industry shares are correlated with gender norms, this positive bias may remain.

Given that no measure of gender norms is observed in the Census, I turn to another dataset, the General Social Survey, to provide some insight as to whether this exclusion restriction for industry shares is plausible. At minimum, for industry shares to be correlated with gender norms in a geographic region, individuals in a certain industry must necessarily be more or less traditional with respect to gender norms. In Appendix C I show that while it is indeed the case that some industries are measurably different in this respect, these differences largely disappear when comparing industries within an educational attainment level. Intuitively, the aggregate differences are mostly driven by the share of higher or lower educated individuals, and it is education that appears to be strongly correlated with norms.

Even after controlling for education, certain industries particularly for men appear to be more traditional than others, such as agriculture for lower-educated men and manufacturing and construction for higher-educated men. In the results I will explore whether these industries contribute significantly to the identifying variation and test the extent to which including such industries matter for the overall results. For women, in contrast, differences essentially disappear upon controlling for education.

Given this mediating role of education, I construct shares for a Bartik instrument based on both industry and education, using two levels: those with a high school degree or less, and those who have attended some college or more. While the educational attainment in a given area may be correlated with norms, conditional on controlling for the overall initial share of educational attainment in the area, the industry shares are presumed to be uncorrelated with gender norms.

Constructing shares in this manner has an added advantage. Within an industry, there are individuals of many different occupations depending on skill level, and wage trends may differ significantly between individuals working in the same industry with different skill levels. For example, while low-educated workers in manufacturing have seen their wages stagnate or decline in recent decades, high-skilled workers' wages in similar industries have grown. A region that has a large share of high-skilled manufacturing would plausibly see different labor market impacts than one with a large share of low-skilled manufacturing.

Considering the industry-education shares as a measure of the jobs available to a given type of individual in their local labor market, industry-level wage shocks will yield different rates of change in the relative wage for otherwise similar couples across different conspumas and over time.

3.4 Estimation

The geographical area that I consider is a unit defined in the U.S. Census IPUMS data, referred to as a Consistent Public Use Microdata Area (hereafter conspuma), of which there are 543. It designates the smallest geographical unit that can be consistently identified between 1980 and 2011.[8] This smaller unit is likely a better proxy for the local labor markets than the state level. I use 3-digit Census Bureau industry codes to characterize industries, with a total of 243 non-missing industries in my sample. Given that I further partition these industries by two levels of educational attainment, there are 486 industry-education shares for men and women respectively.

[8]While counties are otherwise the smallest available unit for public-use microdata within a given year, some counties change borders or merge during this period. Likewise, commuting zones are another geographical unit that is smaller and has been used in other papers employing Bartik instruments, but they are only available in 1980 and 1990.

To construct the instrument, I consider all individuals regardless of marital status who are working and between the ages of 18 and 65. In total, this procedure yields a sample of approximately 20.5 million individuals across all years.

First, I calculate the average log hourly wages by industry, gender, educational level, and year, using the leave-one-out approach to generate a measure of average national wages that leaves out the contribution of each conspuma in turn:

$$\bar{\omega}^g_{iect} = \frac{1}{n}\sum_n \omega^g_{nie-ct}$$

Here, n is an individual subscript, g refers to gender, i refers to industries, e to educational attainment, c to conspuma, and t to year.

Next, I calculate the growth in "national" hourly wages from the preceding period by industry, year, and educational attainment.

$$\Delta^g_{iect} = \frac{\bar{\omega}^g_{iect} - \bar{\omega}^g_{iect-1}}{\bar{\omega}^g_{iect-1}}$$

Finally, I construct the Bartik instruments by using industry shares fixed in 1980 weighted by the growth in wage rates, where the shares are given by θ:

$$\rho^g_{ect} = \sum_i \theta^{1980,g}_{iec} \times \Delta^g_{iect}$$

I ultimately want to estimate the following pair of equations by 2SLS, using the Bartik shocks to instrument for the endogenous wage variables: the relative wage (which in certain specifications includes its square) and the total wages of the couple to distinguish between relative and absolute changes in wages. These endogenous variables are given by the vector Π_{kct}.

First Stage:

$$\Pi_{kct} = \delta_w(\rho_{ct}) + \delta_m(\rho_{jct}) + (X_{kct})\beta + \gamma_t + u_{kct} \tag{4}$$

Second Stage:

$$y_{kct} = (\hat{\Pi}_{ct})\alpha + (X_{kct})\beta + \gamma_t + \epsilon_{kct} \tag{5}$$

Let k denote the individual couple and t the year, where y_{kjt} is the wife's hours, Π_{kct} is the vector of relative wages, its square, and total household wages , X_{kct} represents control variables including dummies for race, Hispanic ethnicity, quadratic terms for age, and years of education of the husband and wife in all specifications,. Time fixed effects are given by μ_t and standard errors are clustered at the conspuma level c.[9]

Additional controls include the share of individuals with some college or more in 1980, the initial period in which industry shares are measured. Conditional on the initial share of educational attainment, the Bartik shocks exploit variation in the share of industries within a given level. In certain specifications I also include the inverse Mills ratio, where dummy variables characterizing the age of the youngest child and whether the couple has no children are used as the exclusion restriction in the probit regression. Given that I focus on women working positive hours, the inclusion of the inverse Mills ratio helps to account for differing propensities to be observed in the labor force.

If the instruments are relevant and the exclusion restriction is valid, α gives the marginal effect of the relative wage on wives' labor supply for those that have a higher relative wage due to labor demand shocks.

For comparison, I also estimate a baseline OLS regression analagous to that carried out for the PSID:

[9] Adão, Kolesár, and Morales (2018) have shown that in certain applications of the Bartik instrument, this type of clustering may yield standard errors that are much too small and lead the null hypothesis to be rejected at much higher rates than expected even when the instrument is known to be irrelevant. The key issue they highlight is that potential outcomes may be correlated across regions with similar distributions of industries, since such regions may also be similar on unobservables. In this setting, there are in fact many units that are exposed to the exact same distributions of industries: all couples in a given conspuma. Thus, clustering at the conspuma level accounts for this aspect of similarity in exposure. However, it is possible that this issue is present beyond the level of the conspuma. Since the structure of my analysis does not allow me to directly implement their procedure for conducting inference, I instead carry out the same placebo test they describe in the paper, described in detail in Appendix E. While rejection rates are slightly higher than expected under this placebo test, the difference is small compared to what Adão et al. (2018) document.

$$h_{kt} = \Pi_{kt}\alpha + X_{kt}\beta + \gamma_t + u_{kt} \tag{6}$$

This regression includes the same baseline controls as those for 2SLS.

4 Results: PSID

To facilitate comparison with the patterns discussed in Sec. 2 I first present graphical evidence using the predicted relative wage from the first-stage regression.

Fig. 5 graphs the mean annual hours of wives and husbands by quantiles of the predicted relative wage and by decade. As in Fig. 4, where annual hours are graphed against the observed relative hourly wage, the level of hours for wives (Panel (a)) appears to be rising across time throughout the distribution of the relative wage, with the largest increases between the 1970s and 1980s. Within each decade, the level of hours rises across quantiles of the relative wage until slightly less than 0.5.

From 1980 onward, to the right of this point the level of hours appears flat, in contrast to the striking decline in hours in Fig. 4 for each decade. In these latter years, women who have a predicted wage greater than their spouse work more than women at lower points in the distribution. The higher-wage women are clustered around an average of about 1800 hours, which is roughly the full-year hours for a full-time 35-hour workweek. For 1970, there is a discernible decline in hours at the middle, though it is much smaller than the decline in hours in Fig. 4 using the observed relative wage. Note that the samples in Fig. 4 and Fig. 5 are identical, with both restricted to couples for which both members of the couple are observed working in two consecutive periods.

For men, it is difficult to observe any meaningful pattern. While there is some variation in hours between about 2000 and 2300 hours per year, across all years and throughout the distribution husbands on average appear to work full-time, whether measured the the observed or predicted relative wage.

This figure provides initial evidence that the relationship between hours and the relative wage may not be inversely U-shaped. Moreover, it casts doubt on the idea that the original pattern is indicative of the breadwinner norm. If the sharp declines in the level of hours with respect to the observed relative wage represented an underlying behavioral response to the breadwinner norm, I would expect it would remain when excluding measurement error as a source of variation in relative wages.

Tables 3 and 5 present the regression results of the OLS and 2SLS respectively. For both tables, the first three columns provide estimates of the linear relationship between the relative wage and hours, while the latter three allow for a quadratic term. The first column includes the full sample and all years. The second column includes the inverse Mills ratio as a control for selection into the labor force. The third column restricts to years prior to 1997, for the reason that after 1997 the PSID is biennial. Thus, when including the post-1997 period, the lags of the wage variables refer to two years earlier, whereas for other years the lags are one year earlier. Finally, it is important to note that in the tables, the relative wage is scaled for ease of reading so that a one-unit increase represents a one percentage-point increase in the relative wage.

The element of interest is the marginal effect of the relative wage across the distribution of relative wages. First, I consider the linear specification. Under OLS, Cols. 1-3 in Table 3 give a negative coefficient on the relative wage, implying that a one percentage-point increase in the relative wage is associated with an decrease of 2 to 3 hours per year. Whereas this value is statistically significant, it is hard to motivate as economically significant.

With a quadratic specification, this marginal effect is given by $\frac{\delta Hours_{wt}}{\delta \pi_{kt}} = \alpha_1 + 2\alpha_2 \pi_{kt}$, where α_1 represents the coefficient on the relative wage and α_2 is the coefficient on its square. Across columns, the quadratic specifications yield similar results, with α_1 ranging from 53.2 to 50.5, and α_2 ranging from -0.61 to -0.59.

The result is easiest to interpret graphically. The marginal effect is shown with 95% confidence intervals in Fig. 6. The first point to note is that the quadratic OLS results mirror the graphical evidence presented in Sec. 2: the marginal effect is positive for values of the relative wage below 50%, crosses zero very near to 50%, and is negative and declining thereafter.

Analysis by 2SLS yields substantially different results. The first stage is shown in Table 4. I present both the standard F-statistics and the Sanderson and Windmeijer (2016) F-statistics, which are more appropriate in a context with multiple endogenous variables. Unsurprisingly given the use of couple-specific lags as instruments, these statistics indicate a very strong first stage. For the linear specification, the SW F-statistic for the relative wage is 610.34, while for the quadratic specification, the corresponding values are 57.12 and 55.03 for the relative wage and its square respectively. The signs of the coefficients are reasonable: the lagged relative wage and its square are positively related to the current-period relative wage; the lagged total household wage is positively related to the current-period total household wage and negatively to the relative wage variables.

The second stage is displayed in Table 5. Whereas when using lagged wages as instruments in Table 5 the corresponding coefficient is positive and much larger in magnitude for the linear specifications, ranging around 26. Similarly to the OLS results, the quadratic specifications yield similar results across columns, with α_1 ranging from 179.2 to 163.1 and α_2 ranging from -1.65 to -1.51.

Fig. 6 graphics the marginal effect for 2SLS in addition to OLS. Both point estimates of the marginal effects are declining in magnitude, and they are both positive for most of the first half of the distribution and negative for the second half. However, the 2SLS estimate is positive and significantly different from zero only for a relative wage less than about 0.4. Between 0.4 and 0.85, the estimate of the marginal effect is not significantly different from zero. For the OLS estimate, which naturally has much smaller standard errors, the marginal effect is positive and significantly different from zero for values lower than about 0.46, and negative and significant after that point. However, the OLS estimate is not significantly different from the 2SLS estimate for values greater than 0.38. For values lower than this, the 2SLS estimate is larger. Even where the 2SLS estimate is negative and statistically significant, it is not particularly meaningful, as it occurs at a value of 0.86 and the highest value of the relative wage in the data is 0.862.

Notably, the 2SLS estimates for the quadratic specification have large standard errors. If lagged relative wages were simply a poor predictor of current relative wages, I would expect the linear specification also to have large standard errors, however this is not the case. The difference in precision across specifications casts doubt on whether the quadratic specification represents the true underlying relationship, and suggests that a linear relationship is a better fit.

5 Results: Census/ACS

Table 6 shows the OLS regression of hours on relative wage (and its square) and the total household wage, analogous to the baseline estimation with the PSID. In the linear specifications in Cols. 1 and 2 (which differ by the inclusion of the inverse Mills ratio), the coefficient on the relative wage is positive rather than negative, taking values of 5.6 and 5.7. Like for the PSID, these coefficients are small in magnitude.

For the quadratic specifications in Cols. 3 and 4, the magnitudes of the coefficients are qualitatively similar to those obtained in Table 3 where the same analysis is carried out for the PSID. The coefficients on the relative wage are 46.13 and 45.01, and for its square they are -0.46 and -0.45 respectively.

The comparison between the PSID and the Census is made clearer in Fig. 7 where the marginal effect of the relative wage from the original PSID regression is graphed alongside that of the Census/ACS. Between approximately 0.2 and 0.45, the confidence intervals overlap. As for the PSID, there is a positive marginal effect for a value lower than 0.5 and a negative value past this point. They do differ in other parts of the distribution, but the similarity is notable given that they come from two different datasets, one cross-sectional and the other a panel, and comprising different years.

Turning to the Bartik analysis, Table 7 displays the first-stage results for both the linear specification and the quadratic specification (when a quadratic term for the relative wage is added). In all cases, the Bartik shocks for men and women respectively as well as their squares are used as instruments.[10] The SW F-stats for a setting with multiple endogenous regressors imply rejection of the null of weak identification for each endogenous regressor. In the linear specification, the SW F-stat is 41.19 for the relative wage (p-value< 0.01), while for the quadratic specification they are 8.00 and 7.40 for the relative wage and its square respectively (both have a p-value< 0.01). Moreover, as both the linear and quadratic specifications are overidentified, I also indicate the Hansen J-stat and associated p-values for each set of equations. In both cases, the Hansen J-stat imply that the overidentifying restrictions are not rejected (the associated p-values are 0.44 for the linear specification and 0.31 for the quadratic).

Table 8 presents the second-stage results. Cols. 1 - 3 present results from the linear specification. The coefficient on the relative wage is positive and highly significant, ranging from 87.11 with the baseline controls to 103.5 with the inverse Mills ratio and controls for 1980 included as well. This is an economically large effect, suggesting that a one-percentage-point increase in the relative wage owing to labor demand shocks leads to the equivalent of two to three weeks of additional full-time work.

The quadratic specification, in Cols. 4 and 5 have a positive coefficient on the relative wage and a negative coefficient for its square, but when 1980 controls are added in Col. 6, controlling for the initial share of each educational group for men and women respectively, both coefficients become positive and insignificant. As in the case of the PSID, this result suggests that the relationship between the relative wage and wives' annual hours are better approximated by a linear specification.

[10] In principle, the two Bartik shocks could be sufficient for a just-identified system in the linear specification where only the relative wage and total household wage are endogenous regressors, however, diagnostic tests suggests the endogenous regressors are only weakly identified in this way. In Appendix D I discuss in more detail the potential issues with using only the two linear Bartik shocks as well as the justification for using their squares as additional instruments. Put simply, if the Bartik shocks are mean independent of the error term, rather than simply uncorrelated, then their squares should also be valid instruments.

5.1 Migration

If the Bartik instruments capture variation in demand in local labor markets, another concern might be that my results are simply capturing migration. For example, it could be that due to favorable labor market conditions, couples with wives highly motivated to work move into the area, increasing the estimated coefficients on the relative wage.

I test for this channel by adding a control for whether the individual lives in the same state as she was born. While this does not map directly to conspumas, those who remain in their state of birth as adults are certainly less mobile than those who have moved states. If more mobile individuals are also those willing to work more, we might expect that the coefficient on remaining in the state of birth would be negative and that the coefficient on the relative wage would decline when the control is included.

The results of this exercise are presented in Table 9. Col. 1 reproduces the result from Col. 3 of Table 8 without the state of birth dummy for comparison. In Col. 2, inclusion of the state of birth dummy variable leaves the coefficient on the relative wage essentially unchanged, and the coefficient on the state of birth variable itself is not significantly different from zero. This result suggests that more mobile women are not necessarily those that work more. Col. 3 and Col. 4 corresponds to the results from Col. 3, Table 8, comparing with and without the state of birth variable. Again, the results are similar with or without the control, and the coefficient on the state of birth dummy is not significant.

Given these results, migration patterns do not appear to be driving the original positive relationship between the relative wage and hours.

5.2 Heterogeneity by Marital Match

Another potential issue is that the average effect across the population may be masking heterogenous effects. As such, it may be that certain subgroups–perhaps those more likely to care about gender norms–do have a negative response to increases in the relative wage. To provide some insight on this question, I estimate the main specification for four different "couple-types" based on combinations of the two levels of education used throughout the analysis.

Table 10 displays the results. Col. 1 includes couples where both have a high school education or less. The coefficient on the relative wage is 38.93, and highly statistically significant. The result is similar in Col. 2, which includes couples where the wife is less-educated and the husband is more-educated, with a coefficient of 34.43. In Col. 3, which includes couples that both have at least some college, the coefficient on the relative wage remains highly significant and is larger in magnitude, at 66.99. Finally, in Col. 4, representing couples where the wife is highly-educated and the husband has the lower level, magnitude is smaller but still significant, at 43.69.

These separate regressions do not provide support for the hypothesis that some groups indeed might have a negative effect on hours in response to increases in the relative wage, as across all groups the effect of the relative wage is positive.

5.3 Assessing the Identifying Variation

In the recent papers studying Bartik or "shift-share" instruments, one of the key contributions has been to define methods to unpack and assess the identifying variation contained in the instrument. Given its "black-box" construction, the practitioner does not know *a priori* which industries make the Bartik instrument relevant.

While my setting precludes the methods proposed by Borusyak, Hull, and Jaravel (2018),[11] I can calculate the Rotemberg weights as detailed in Goldsmith-Pinkham et al. (2018) (hereafter GPSS).

A key practical concern that GPSS demonstrate is that in many applications a very small number of industries receive disproportionately high weights. In the

[11] Their approach hinges on rewriting the Bartik analysis at the industry level, which in turn relies on the standard setup where outcomes and treatments are defined at the same "location" level so that location-level variables can be reweighted by their industry exposure. Since my instruments are defined at the location level but the outcomes at the individual household level, my analysis cannot be rewritten in this way.

applications they consider, the five highest-weight industries often receive a third to a half of the weight, even with hundreds of industries included. When industries have disproportionate weight, one can think of the Bartik IV strategy as essentially comparing across locations based on variation in exposure to these few industries. Thus, it is important to consider whether these industries are potentially problematic with respect to the exclusion restriction.

Fig. 8 shows the cumulative fraction of the total positive weight by the rank of the Rotemberg weight for a given industry. On the same plots, a second line graphs the cumulative share of the weights that owes to industries defined by the higher educational group. There are two points to note. First, a minority of industries are disproportionately important, but perhaps not to the extent of some of the applications explored in GPSS.

The top five industries for the female instrument account for only a fifth of of the positive weight due to the Rotemberg weights, while the corresponding share for the top five industries for males is slightly less than one third. In comparison, the 25th to 50th-ranked industries for females also account for roughly a fifth of the overall weight. Table 11 presents details for the top ten-weighted industries, including the name of the industry, the educational level with which it is associated, the average share and its standard deviation, the average growth weight across years and industries, and the Rotemberg weights themselves.

Second, while the share of industries driving the positive weight for the female instrument is nearly equally from high- and low-education industries over all industries, for the male instrument it is overwhelmingly due to the low-education industries. This pattern is marked for the first ten industries, of which all but two are a low-education industry.

Next, I turn to examining the top industries themselves. For the female instrument, the top ten industries seem quite varied by sector, with a couple of manufacturing-based industries, a couple in professional services, and a couple in other types of services. None of these industries jump out as potentially problematic for the exclusion restriction.

However, for the male instrument, the top four industries industries include "Agriculture" and "All construction" (both associated with high school or less) and both educational levels of "Oil and gas extraction." As discussed in Appendix C, where data on gender norms is analyzed to determine whether there are meaningful differences across industries, agriculture for lower-educated men was the only industry that distinguished itself as significantly more conservative than others, whether or not one conditioned on education. Without conditioning on education, those working in construction were also significantly more conservative than others. Finally, although the GSS data did not allow me to examine "Oil and gas extraction" there are journalistic accounts that "oil boomtowns" are extremely unusual with respect to their population (Gottesdiener, 2014). They tend to have high levels of young, single, opportunistic men lured by the promise of high-paying if potentially temporary jobs. These areas also usually only "boom" for a limited time period, and then collapse.

That these represent the highest-weighted male industries is potentially problematic, as there are clear possible correlations with gender norms for each one. The concern then is that the identifying variation compares across couples living in conspumas with higher or lower shares of industries that may be associated with traditional gender norms. This type of concern may extend more broadly than to the first three industries, especially given that the male instrument places much more weight on lower-education industries than does the female instrument. Could this be a reason that I consistently find a positive relationship between the relative wage and wives' hours?

To address this concern, I reestimate the IV model, but rather than using both of the full Bartik instruments, I will use industry-education-level shocks for men and women separately and compare the results. I estimate specifications using shocks from the top five, ten, and fifty industry-education categories according to their Rotemberg weights, and control for the total share that these categories represent in a given conspuma in 1980.[12]

Table 12 presents the results. Using the top five industry-education categories for women, the coefficient on the relative wage is positive (18.96) but not significantly different from zero, although the F-stat indicates a strong first stage (29.56). The overidentifying restrictions are not rejected. When using the top ten and 50 categories, the coefficient becomes larger in magnitude (28.97 and 21.52) and the standard error declines, resulting in a statistically significant effect. However, neither of the latter coefficients are statistically different from the first.

Each of the coefficients on the relative wage using the male shocks is positive and significant, ranging from 23.8 to 30.53. The magnitude of these estimates are similar to those for the female-only shocks.

The magnitude of the estimates are of course substantially smaller compared to when both full Bartik instruments are used. This may perhaps owe to the greater range of variation that is exploited when both male and female shocks are included.

Lastly, the overidentifying restrictions are rejected in all but the first specification for the female shocks. As GPSS note themselves, it is common for these tests to reject in heavily overidentified settings. However, in this case, the fact that the separate instruments for men and women yield similar results to using the full Bartik instruments provides some additional evidence that the casual effect of the relative wage is positive. To the extent that norms might be correlated with the shares of the top-weighted male shocks, this potentially confounding factor at least appears to have little effect on the overall results.

6 Discussion

Given the inverse U-shaped relationship I initially found between hours and the relative wage, I allowed for the possibility that the true relationship might be quadratic,

[12] Borusyak et al. (2018) show when using a Bartik shock where the shares do not sum to one, failing to control for the sum of the shares or its complement can introduce bias by implicitly putting a zero-shock on the industries that are unaccounted for, which varies in importance across locations.

with hours declining as the wife's wage exceeded the husband's. However, this apparent relationship may owe to measurement error. When instrumenting for the relative wage and total household wage in two different ways neither approach provides support for a quadratic relationship. Whether these strategies are valid depends on whether the instruments have a relationship to the relative wage and whether they are plausibly independent of the measurement error in wages.

It is hard to find stronger instruments for wages than using lagged wages for the same individuals, as in the analysis of panel data. However, a concern here remains that measurement error may be serially correlated. As such, the analysis using the Bartik instrument is reassuring, since the Bartik instruments are plausibly independent of measurement error in wages, and though they have a weaker first stage than lagged wages they are nevertheless strong by conventional standards.

Finally, even though the relationship between wives' hours and the relative wage is positive in the PSID analysis, these estimates might not be representative of the causal effect of the relative wage, since I use individuals' observed wages rather than their potential wage offers. However, when using the Bartik instruments to capture the effect on hours of an increase in the relative wage owing to changes in labor demand, the relationship is still positive. This result suggests that the causal effect of the relative wage on wives' hours is positive, contrary to one potential mechanism for the breadwinner norm.

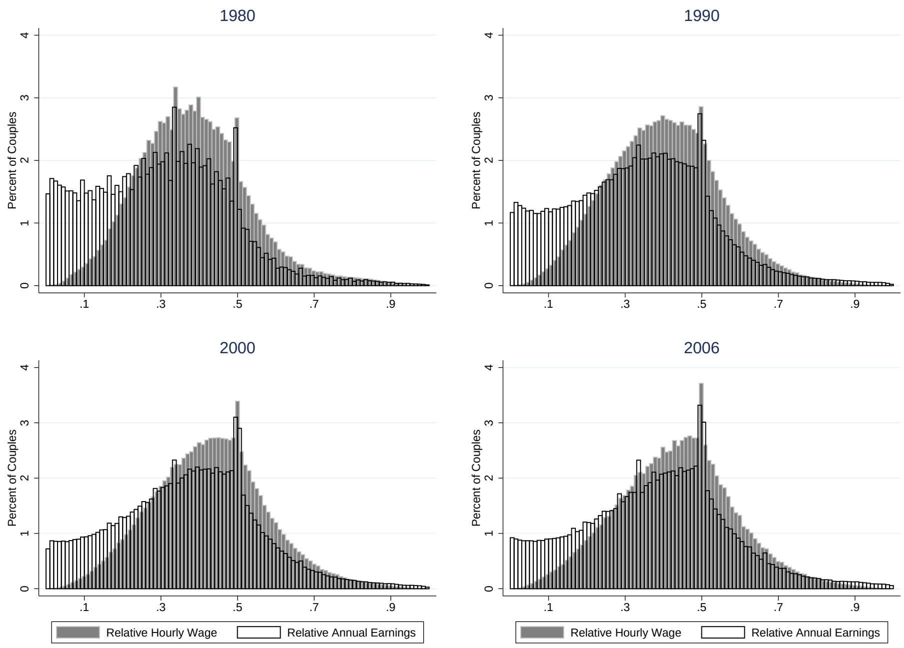
7 Conclusion

This paper contributes to the understanding of the mechanisms by which gender norms do–or don't–operate on married women's behavior. I revisit the notion of the breadwinner norm and how it may affect women's labor supply decisions, conditional on being married. Focusing on the intensive margin, I find no evidence that women work less as their probability of outearning their husbands increases. The inverse U-shaped relationship that appeared consistent with the breadwinner norm story disappears when accounting for measurement error. When considering the causal effect of an increase in the relative wage, the effect in linear specifications is consistently positive, and I find little evidence that the true relationship is in fact quadratic. I have also shown that these results are robust when accounting for alternative explanations, estimating the effect for heterogenous subgroups, and considering possible confounding factors.

While the breadwinner norm may not affect household labor decisions within marriage, this is not to say the norm may not matter at all. BKP also provide evidence that the breadwinner norm may affect entry into marriage, and this would be perfectly consistent with my finding of no effect conditional on marriage. Indeed, the lack of an effect within marriage suggests that perhaps greater attention should be given to the ways other gender norms also may affect selection into marriage.

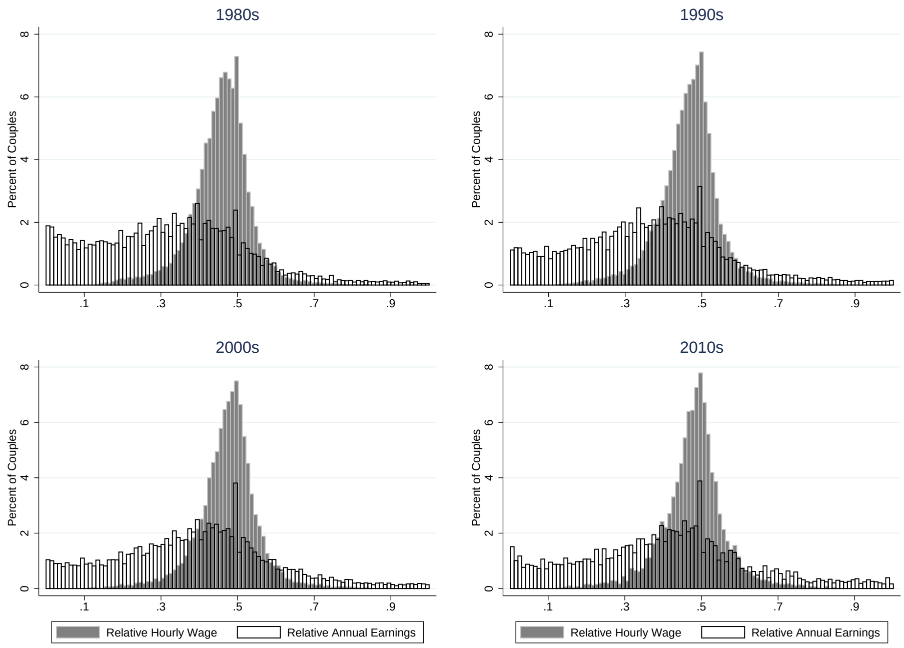
8 Figures

Figure 1: Census and American Community Survey: Distribution of Within-Couple Relative Income and Wages



The data is drawn from U.S. Census waves 1980, 1990, and 2000, and the ACS waves 2005-2007 (which are pooled to be of similar size to the other waves). The sample includes couples where both members are between the ages of 18 and 65, work positive hours, and are not self-employed. The relative wage (income) is calculated by dividing the wife's hourly wage (annual wage income) by the sum of the wife and husband's hourly wages (annual wage income).

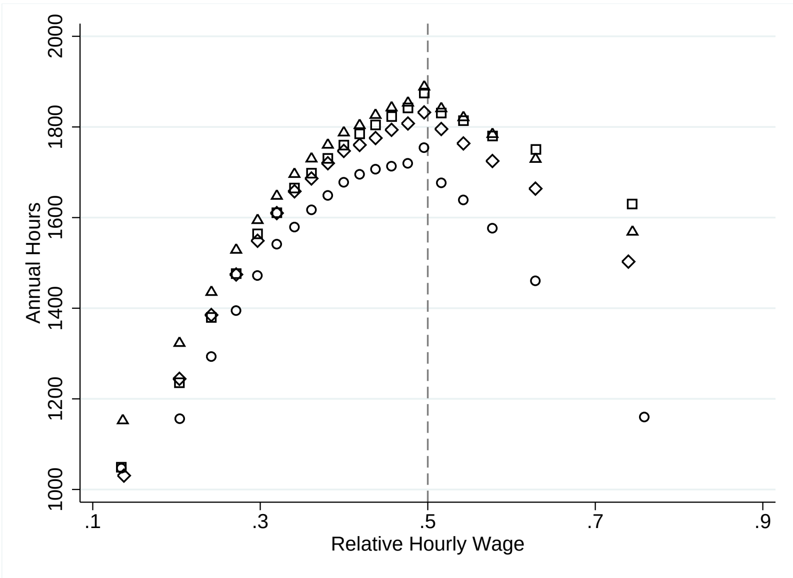
Figure 2: PSID: Distribution of Within-Couple Relative Income and Wages



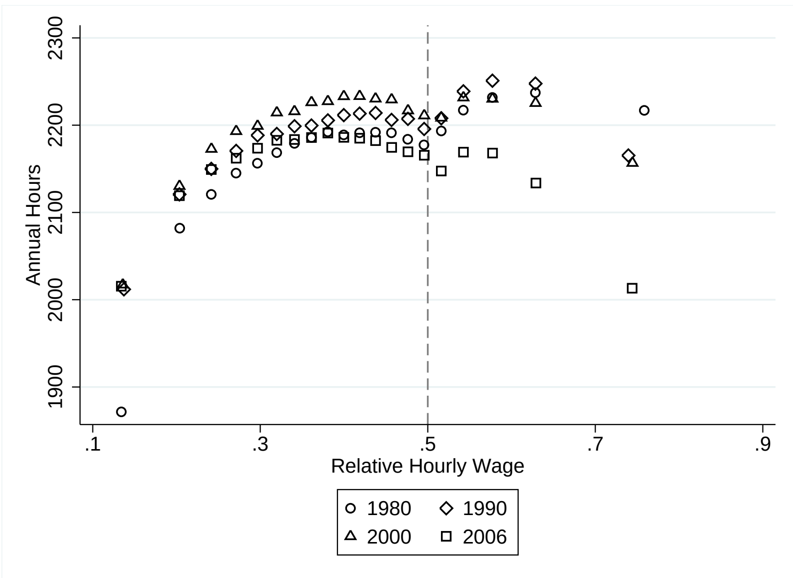
The data is taken from the Panel Study of Income Dynamics, between the years 1969 to 2013 (noting that from the year 1997, the survey is conducted biennially.) The sample includes couples where both members are between the ages of 18 and 65, work positive hours, and are not self-employed. The relative wage (income) is calculated by dividing the wife's hourly wage (annual wage income) by the sum of the wife and husband's hourly wages (annual wage income).

Figure 3: Census/ACS: Annual Hours of Married Individuals by Quantile

(a) Wives

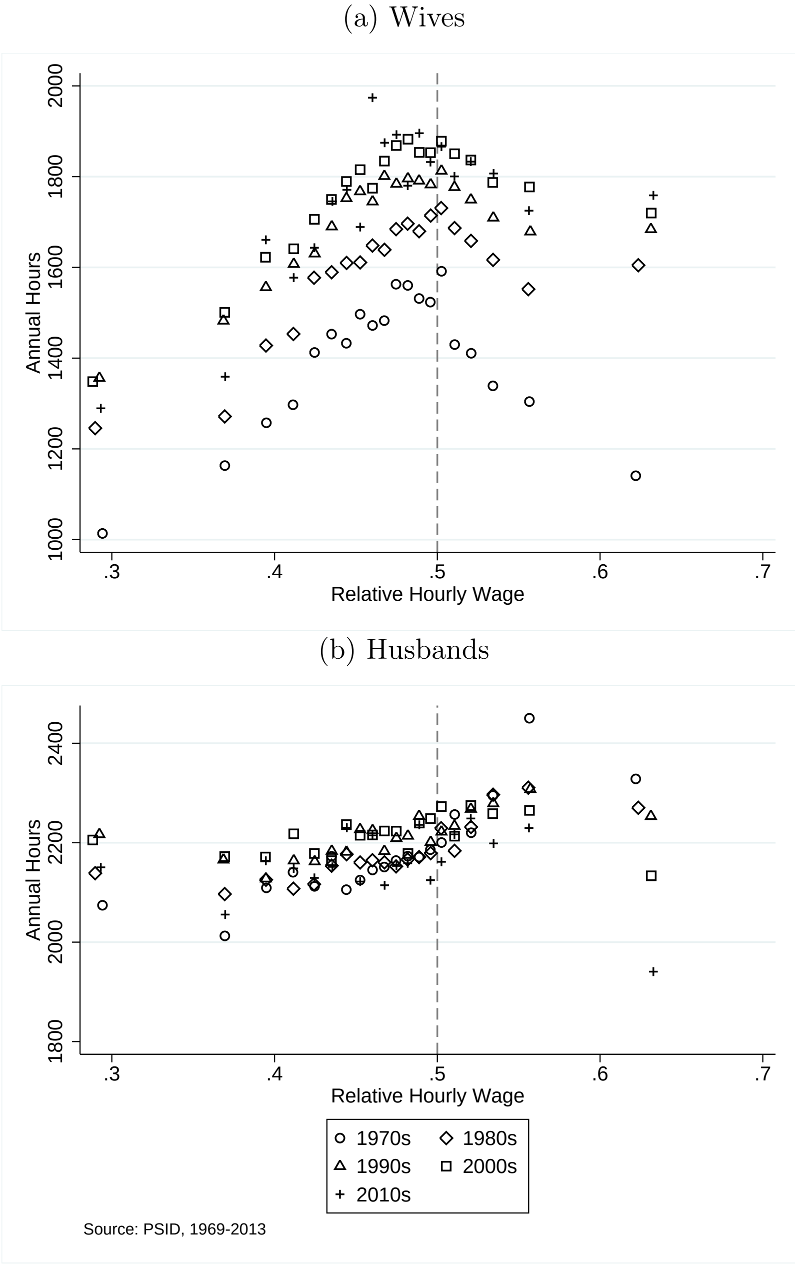


(b) Husbands



Panel (a) and (b) graph the mean annual hours worked by wives and husbands respectively by quantiles of the relative wage for a given year. The data used are from decennial waves of the Census from 1980 to 2000, and the single-year waves froom 2005-2007 are pooled, denoted by 2006. The sample includes individuals between the ages of 18 and 65 in couples where both spouses work positive hours and neither have imputed earnings. Hourly wages are calculated by dividing annual income by the product of usual weekly hours and weeks worked in the previous year. The relative hourly wage is the wife's hourly wage as a fraction of the sum of both spouses' hourly wages.

Figure 4: PSID: Annual Hours of Married Individuals by Quantile of Relative Wage

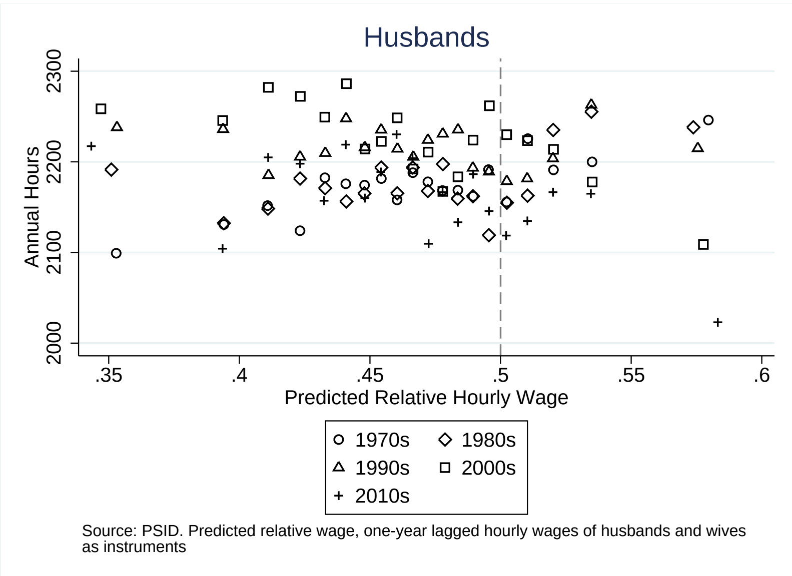


Panel (a) and (b) graph the mean annual hours worked by wives and husbands respectively by quantiles of the relative hourly wage for a given decade. The data used are from the Panel Study of Income Dynamics, 1969-2013, including all couple-year observations for married couples. The sample includes individuals between the ages of 18 and 65 in couples where both spouses work positive hours. Hourly wages are calculated by dividing annual income by annual hours. The relative hourly wage is the wife's hourly wage as a fraction of the sum of both spouses' hourly wages.

Figure 5: Annual Hours of Married Individuals by Quantile of Predicted Relative Wage

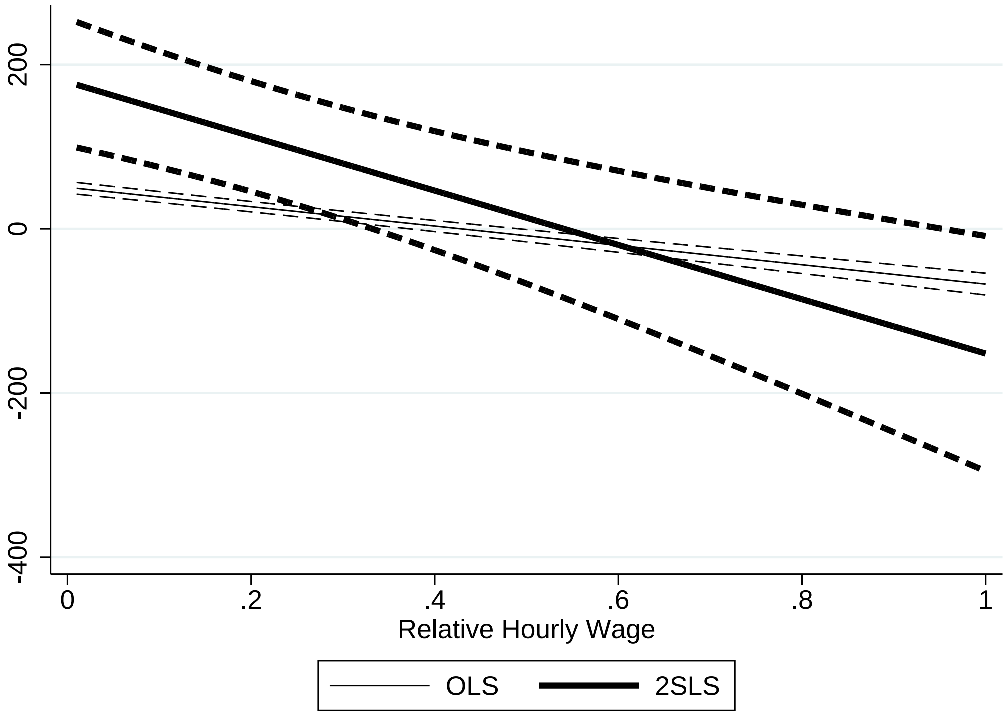
(a) Wives

(b) Husbands



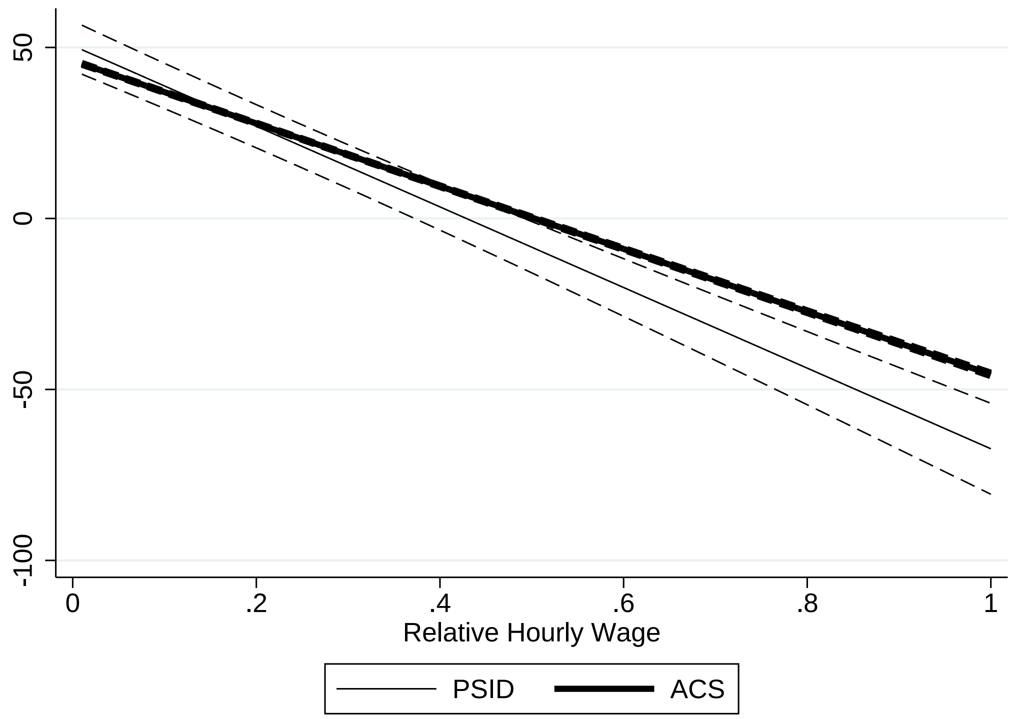
Panel (a) and (b) graph the mean annual hours worked by wives and husbands respectively by quantiles of the predicted relative wage for a given decade. The data used are from the Panel Study of Income Dynamics, 1969-2013, including all couple-year observations for married couples. The sample includes individuals between the ages of 18 and 65 in couples where both spouses work positive hours. Predicted relative hourly wages are estimated from a regression of the relative wage in time t on the relative hourly wage in time $t-1$, the total household wage in time $t-1$, as well as quadratics in the age of each spouse, the number of children in the household, year fixed effects, and couple fixed effects.

Figure 6: Marginal Effect of the Relative Wage (PSID))



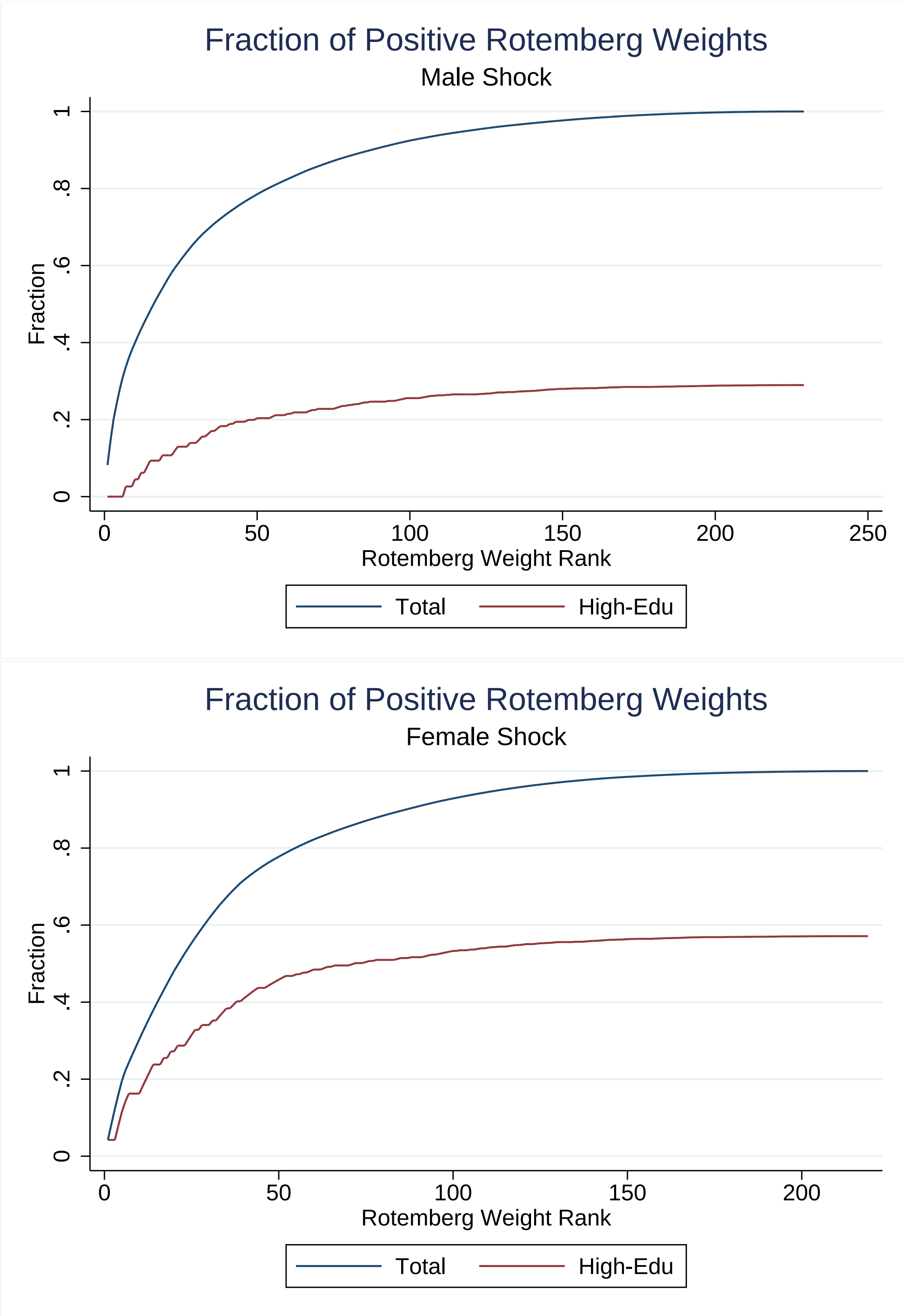
The graph depicts the marginal effect of the relative wage on married women's annual hours conditional on the level of the relative wage, for estimation that does and does not take into account measurement error in wages (OLS and 2SLS respectively). Dotted lines give 95% confidence intervals. For OLS, the marginal effects are based on parameters estimated in Col. 6 of Table 3. For 2SLS, the marginal effects are based on parameters estimated in Col. 6 of Table 5.

Figure 7: Marginal Effect of the Relative Wage (PSID and ACS/Census)



The graph depicts the marginal effect of the relative wage on married women's annual hours conditional on the level of the relative wage, comparing OLS estimates across the PSID and the ACS/Census samples (e.g. not addressing measurement error). Dotted lines give 95% confidence intervals. For the PSID, the marginal effects are based on parameters estimated in Col. 6 of Table 3. For the ACS/Census, the marginal effects are based on parameters estimated in Col. 6 of Table 6.

Figure 8: Cumulative distributions of Rotemberg weights



9 Tables

Table 1: ACS/Census Summary Statistics

(a) Individual Characteristics

	Wives		Husbands	
	Mean	St. Dev.	Mean	St. Dev.
Demographics				
White	0.86	(0.35)	0.86	(0.35)
Black	0.06	(0.24)	0.07	(0.25)
Other Races	0.08	(0.27)	0.07	(0.26)
Age	40.06	(9.97)	42.16	(10.25)
Hispanic	0.07	(0.26)	0.07	(0.26)
Education				
<HS	0.07	(0.25)	0.09	(0.29)
HS Grad	0.40	(0.49)	0.38	(0.48)
SomeColl	0.26	(0.44)	0.25	(0.43)
College	0.27	(0.44)	0.28	(0.45)
Income and Work				
Hours (Annual)	1680	(704)	2195	(587)
Wage Income (Annual)	28939	(23332)	51229	(33757)
Log Wage (Hourly)	2.64	(0.62)	3.00	(0.58)
Bartik Shock	0.06	(0.04)	0.01	(0.04)
Observations	2663078		2663078	

(b) Household Characteristics

No. of Children	1.25	(1.15)
Relative Income	0.36	(0.19)
Total HH Income	87065.38	(48993.01)
Relative Wage	42.06	(15.24)
Total HH Wage	5.65	(0.96)
Observations	2663078	

The table presents summary statistics for married couples' individual and household characteristics from Census waves 1990 to 2000 and ACS waves 2005-2007. Couples are included if they are between the ages of 18 and 65 and both are working positive hours, and neither are self-employed. Income and wages are given in 2007 dollars.

Table 2: PSID Summary Statistics

(a) Individual Characteristics

	Wives		Husbands	
	Mean	St. Dev.	Mean	St. Dev.
Age	37.76	(10.34)	40.06	(10.67)
Hours (Annual)	1627.37	(661.76)	2187.78	(624.56)
Annual Wage Income	22456.01	(17100.91)	41690.82	(26010.88)
Hourly Log Wage	2.45	(0.59)	2.81	(0.57)
Observations	51885		51885	

(b) Household Characteristics

No. of Children	1.27	(1.25)
Relative Income	0.35	(0.18)
Total HH Income	72226.40	(41707.31)
Relative Wage	46.48	(7.18)
Total HH Wage	5.26	(0.94)
Observations	51885	

The table presents summary statistics for married couples' individual and household characteristics from the PSID between 1969-2013. Couples are included if they are between the ages of 18 and 65, both are working positive hours, and neither are self-employed. Income and wages are given in 1999 dollars.

Table 3: Observed Relative Wage and Hours (PSID)

Dependent Variable:	Linear			Quadratic		
Wives' Annual Hours	(1)	(2)	(3)	(4)	(5)	(6)
$RelWage_t$	-2.643	-2.686	-3.287	53.232	53.213	50.540
	(0.548)	(0.548)	(0.645)	(3.141)	(3.141)	(3.693)
$RelWage_t^2$				-0.605	-0.605	-0.589
				(0.033)	(0.033)	(0.039)
Observations	51,885	51,885	38,649	51,885	51,885	38,649
R-squared	0.079	0.081	0.083	0.091	0.093	0.094
Number of famid	6,882	6,882	5,284	6,882	6,882	5,284
Pre-1997	NO	NO	NO	NO	NO	NO
Inverse Mills Ratio	NO	YES	YES	NO	YES	YES

Robust standard errors in parentheses

Results from OLS regression of wives' annual hours worked on the relative hourly wage (and its square in Cols. 4-6). The baseline sample includes couples where both members are between 18 and 65 and both work positive hours. All regressions include a quadratic in husband and wife's age, the total household wage (the sum of their hourly wages), year fixed effects, the number of children, and couple fixed effects. The inverse Mills ratio is calculated from predictions based on single-year probit regressions of wives' labor force participation on dummy variables for the age of the youngest child (infant, toddler, preschool, primary-school, high-school), the number of children, quadratics in the husband and wife's age, dummies for the husband's race (the wife's race is not recorded separately until the 1980s), and dummies for educational attainment (high school dropout, high school graduate, some college, and college graduate) for both wife and husband.

Table 4: PSID: First-Stage Results

	Linear		Quadratic		
VARIABLES	$RelWage_t$	$TotHHWage_t$	$RelWage_t$	$RelWage_t^2$	$TotHHWage_t$
$RelWage_{t-1}$	0.170	-0.002	0.268	17.58	-0.0150
	(0.007)	(0.001)	(0.0423)	(3.974)	(0.00358)
$RelWage_{t-1}^2$			-0.00107	-0.0201	0.000143
			(0.000457)	(0.0449)	(3.87e-05)
$TotWage_{t-1}$	-0.145	0.244	-0.221	-23.25	0.254
	(0.063)	(0.007)	(0.0663)	(6.169)	(0.00708)
Observations	51,885	51,885	51,885	51,885	51,885
Number of famid	6,882	6,882	6,882	6,882	6,882
F-stat	305.15	702.6	206.83	214.81	473.67
F p-val	0.00	0	0.00	0.00	0.00
SW F-Stat	610.34	1406	57.12	55.03	637.49
SW p-val	0.00	0	0.00	0.00	0.00

Robust standard errors in parentheses

Results from the first-stage regressions using the lagged relative wage (and its square where relevant) and lagged total household wage as instruments for the relative wage (and its square) and total household wage. The first stages correspond to the second-stage regressions in Table 5 in Cols. 2 and 5 respectively. The regressions include a quadratic in husband and wife's age, year fixed effects, the number of children, and couple fixed effects, and the inverse Mills Ratio. SW refers to F-statistics calculated based on Sanderson and Windmeijer (2016).

Table 5: Relative Wage and Wives' Annual Hours: IV using Lagged Wages (PSID)

Dependent Variable:	Linear			Quadratic		
Wives' Annual Hours	(1)	(2)	(3)	(4)	(5)	(6)
$RelWage_t$	26.887	26.758	26.132	179.232	178.847	163.103
	(3.107)	(3.106)	(3.646)	(39.504)	(39.439)	(45.514)
$RelWage_t^2$				-1.657	-1.654	-1.508
				(0.422)	(0.421)	(0.492)
Observations	51,885	51,885	38,649	51,885	51,885	38,649
Number of famid	6,882	6,882	5,284	6,882	6,882	5,284
Pre-1997	NO	NO	YES	NO	NO	YES
Selection Corrected	NO	YES	YES	NO	YES	YES

Robust standard errors in parentheses

Results from 2SLS estimation of wives' annual hours worked on the relative hourly wage (and its square in Cols. 4-6). The baseline sample includes couples where both members are between 18 and 65 and both work positive hours. All regressions include a quadratic in husband and wife's age, year fixed effects, the number of children, and couple fixed effects. The inverse Mills ratio is also included as a control in the "Selection Corrected" regressions.

Table 6: Observed Relative Wage and Hours (ACS/Census)

Dependent Variable:	Linear		Quadratic	
Wives' Annual Hours	(1)	(2)	(3)	(4)
$RelWage$	5.643	5.713	46.130	45.008
	(0.034)	(0.033)	(0.136)	(0.135)
$RelWage^2$			-0.459	-0.445
			(0.002)	(0.001)
Observations	2,663,078	2,663,078	2,663,078	2,663,078
R-squared	0.036	0.067	0.080	0.108
Selection Corrected	NO	YES	NO	YES

Robust standard errors in parentheses

Results from OLS regression of wives' annual hours worked on the relative hourly wage (and its square in Cols. 4-6) with robust standard errors. The baseline sample includes couples where both members are between 18 and 65 and both work positive hours. All regressions include a quadratic in husband and wife's age, the total household wage (the sum of their hourly wages), year fixed effects, dummies for educational attainment (high school dropout, high school graduate, some college, college graduate). The inverse Mills ratio is included in the "Selection Corrected" regressions.

Table 7: Bartik IV Estimates: First-Stage Results (ACS/Census)

	Linear		Quadratic		
VARIABLES	$RelWage$	$TotHHWage$	$RelWage$	$RelWage^2$	$TotHHWage$
$Bartik_w$	-3.582	10.656	-3.582	-668.058	10.656
	(12.727)	(1.847)	(12.727)	(1,033.259)	(1.847)
$Bartik_m$	-0.253	-0.002	-0.253	166.570	-0.002
	(8.284)	(1.265)	(8.284)	(667.592)	(1.265)
$Bartik_w^2$	360.333	-48.164	360.333	26,091.056	-48.164
	(56.484)	(9.058)	(56.484)	(4,755.775)	(9.058)
$Bartik_m^2$	-146.686	-1.981	-146.686	-7,019.254	-1.981
	(49.477)	(8.563)	(49.477)	(3,905.410)	(8.563)
Observations	2,663,078	2,663,078	2,663,078	2,663,078	2,663,078
F-stat	30.59	19.21	30.59	16.74	19.21
F p-val	0.00	0.00	0.00	0.00	0.00
SW F-Stat	41.19	27.56	8.00	7.40	8.59
SW p-val	0.00	0.00	0.00	0.00	0.00
Hansen J	1.66		1.03		
Hansen p-val	0.44		0.31		

Results using Census/ACS data from the first-stage regressions using Bartik shocks as instruments for the relative wage (and its square) and total household wage. The first stages correspond to the second-stage regressions in Table 8 in Cols. 3 and 6 respectively. Standard errors are clustered at the conspuma level. SW refers to F-statistics calculated based on Sanderson and Windmeijer (2016).

Table 8: Relative Wage and Wives' Annual Hours: Bartik IV Estimates (ACS/Census)

Dependent Variable:	Linear			Quadratic		
Wife's Annual Hours	(1)	(2)	(3)	(4)	(5)	(6)
$RelWage$	52.982	50.693	50.939	107.628	87.611	88.477
	(8.236)	(7.776)	(7.863)	(36.148)	(36.942)	(36.570)
$RelWage^2$				-0.759	-0.511	-0.520
				(0.496)	(0.510)	(0.504)
Observations	2,663,078	2,663,078	2,663,078	2,663,078	2,663,078	2,663,078
Local Controls	NO	NO	YES	NO	NO	YES
Selection Corrected	NO	YES	YES	NO	YES	YES

Robust standard errors in parentheses

Results from 2SLS regression of wives' annual hours worked on the relative hourly wage (and its square in Cols. 4-6). The baseline sample includes couples where both members are between 18 and 65 and both work positive hours. The Bartik shocks and their squares are the instruments in each equation. The relative wage, (and its square where relevant) and the total household wage are treated as endogenous. All regressions include a quadratic in husband and wife's age, year fixed effects, dummies for educational attainment (high school dropout, high school graduate, some college, college graduate) and dummies for race and Hispanic ethnicity. Local controls are the share of individuals in a conspuma with greater than a high school education in 1980. The inverse Mills ratio is included as a control in regresisons marked "Selection Corrected."

Table 9: Relative Wage and Wife's Annual Hours: Migration

Dependent Variable:	Linear		Quadratic	
Wife's Annual Hours	(1)	(2)	(3)	(4)
$RelWage$	50.939	51.115	88.477	88.305
	(7.863)	(7.877)	(36.570)	(38.613)
$RelWage^2$			-0.520	-0.525
			(0.504)	(0.541)
Born in state		2.350		-4.589
		(3.733)		(7.477)
Observations	2,663,078	2,663,078	2,663,078	2,663,078

Robust standard errors in parentheses

Results from 2SLS regression of wives' annual hours worked on the relative hourly wage. Cols. 2 and 4 differ from Cols. 1 and 3 by the inclusion of a dummy variable indicating whether the wife was born in the same state as she is currently living. All other aspects of the regression are the same as described for Col. 3 and Col. 6 of Table 8, for the linear and quadratic specifications respectively.

Table 10: Relative Wage and Wife's Hours: Bartik IV Heterogeneity

Dependent Variable:				
Wife's Hours	(1)	(2)	(3)	(4)
$RelWage$	38.936	34.429	66.986	43.686
	(6.322)	(5.256)	(19.489)	(11.423)
Observations	874,968	363,683	1,049,648	374,779
Edu_w	Low	Low	High	High
Edu_m	Low	High	High	Low

Robust standard errors in parentheses

Results from 2SLS regression of wives' annual hours worked on the relative hourly wage, for four combinations of high- or low-educated wives and husbands (highly educated refers to some college or more, while low-educated refers to high school or less). The baseline sample includes couples where both members are between 18 and 65 and both work positive hours. The relative wage and the total household wage are treated as endogenous and Bartik shocks for men and women are used to instrument. All regressions include a quadratic in husband and wife's age, year fixed effects, years of education, dummies for race and Hispanic ethnicity, the inverse Mills ratio, and the initial share of highly educated in a given conspuma in 1980.

Table 11: Summary statistics for top-weighted industries, by Rotemberg weights

(a) Women

Industry	Edu. Level	Mean Share	St.Dev. Share	Mean Growth	Rot. Weight
Yarn, thread, and fabric mills	0	0.87	2.64	0.21	0.06
Motor vehicles and motor vehicle equipment	0	0.68	1.39	0.15	0.12
Banking	0	1.55	0.51	0.12	0.06
Banking	1	0.66	0.33	0.20	0.13
Insurance	0	1.32	1.01	0.15	0.13
Insurance	1	0.61	0.47	0.20	0.12
Hotels and motels	1	0.20	0.26	0.26	0.07
Hospitals	0	3.21	1.12	0.08	0.07
Legal services	1	0.30	0.22	0.24	0.11
General government, n.e.c.	1	0.43	0.39	0.17	0.08

(b) Men

Industry	Edu. Level	Mean Share	St.Dev. Share	Mean Growth	Rot. Weight
Agricultural production, crops	0	1.32	2.06	0.09	2.01
Oil and gas extraction	0	0.81	1.93	0.18	1.55
Oil and gas extraction	1	0.41	0.83	0.21	0.65
All construction	0	8.44	2.72	-0.01	1.37
Meat products	0	0.50	0.85	0.06	0.98
Apparel and accessories, except knit	0	0.36	0.50	0.13	0.91
Industrial and miscellaneous chemicals	0	0.61	1.08	0.11	0.51
Industrial and miscellaneous chemicals	1	0.43	0.84	0.19	0.45
Grocery stores	0	1.13	0.36	-0.04	0.58
Eating and drinking places	0	1.27	0.61	-0.01	0.79

These tables list the top industry-education categories for the Bartik instrument for women and men respectively. Educational level is denoted by 0 (high school degree or less) and 1 (some college and more). Wage growth rates and shares are multipled by 100 for legibility.

Table 12: Relative Wage and Wife's Hours: Overidentified Estimates

Dependent Variable:	Female Shocks			Male Shocks		
Wife's Annual Hours	(1)	(2)	(3)	(4)	(5)	(6)
RelWage	24.594	25.593	18.146	28.451	27.175	26.716
	(9.199)	(5.665)	(4.369)	(8.205)	(3.960)	(3.756)
Observations	2,663,078	2,663,078	2,663,078	2,663,078	2,663,078	2,663,078
SW: RelWage	10.50	18.39	9.44	20.10	37.76	13.13
SW p-val	0.00	0.00	0.00	0.00	0.00	0.00
SW: TotHHWage	10.47	25.52	21.49	12.35	31.02	19.40
SW p-val	0.00	0.00	0.00	0.00	0.00	0.00
Hansen's J-Stat	27.24	33.33	140.56	16.02	25.31	156.28
p-value	0.00	0.00	0.00	0.00	0.00	0.00

Robust standard errors in parentheses

Results from 2SLS regression of wives' annual hours worked on the relative hourly wage. Cols. 1-3 use the top five, 10, and 50 individual Bartik shocks for women as instruments, while Cols. 4-6 use the top five, 10, and 50 individual Bartik shocks for men as instruments, defined by Rotemberg weight ranks. The relative wage and the total household wage are treated as endogenous and Bartik shocks for men and women are used to instrument. All regressions include a quadratic in husband and wife's age, year fixed effects, years of education, dummies for race and Hispanic ethnicity, the inverse Mills ratio, and the initial share of highly educated in a given conspuma in 1980. SW F-statistics and Hansen J statistics refer to the first stage.

10 References

(2018). Panel study of income dynamics [dataset]. *Ann Arbor, MI: University of Michigan*.

Adão, R., M. Kolesár, and E. Morales (2018). Shift-share designs: Theory and inference. Technical report.

Aizer, A. (2010). The gender wage gap and domestic violence. *American Economic Review 100*(4), 1847–59.

Avdic, D. and A. Karimi (2018). Modern family? Paternity leave and marital stability. *American Economic Journal: Applied Economics 10*(4), 283–307.

Bardasi, E. and J. C. Gornick (2008). Working for less? Women's part-time wage penalties across countries. *Feminist economics 14*(1), 37–72.

Bartik, T. J. (1991). *Who benefits from state and local economic development policies?* W.E. Upjohn Institute for Employment Research.

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 (2018). Is there a male breadwinner norm? The hazards of inferring preferences from marriage market outcomes. Technical report, National Bureau of Economic Research.

Blau, F. D. and L. M. Kahn (2007). Changes in the labor supply behavior of married women: 1980–2000. *Journal of Labor Economics 25*(3), 393–438.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Borjas, G. J. (1980). The relationship between wages and weekly hours of work: The role of division bias. *The Journal of Human Resources 15*(3), 409–423.

Borusyak, K., P. Hull, and X. Jaravel (2018). Quasi-experimental shift-share research designs. Technical report, National Bureau of Economic Research.

Cochard, F., H. Couprie, and A. Hopfensitz (2018). What if women earned more than their spouses? An experimental investigation of work-division in couples. *Experimental Economics 21*(1), 50–71.

Dieterle, S. G. and A. Snell (2016). A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument. *Labour Economics 42*, 76–86.

Fernández, R. (2013). Cultural change as learning: The evolution of female labor force participation over a century. *American Economic Review 103*(1), 472–500.

Folk, O. and J. Rickne (2019). All the single ladies: Job promotions and the durability of marriage. *American Economic Journal: Applied Economics Forthcoming.*

Fortin, N. M. (2005). Gender role attitudes and the labour-market outcomes of women across OECD countries. *Oxford Review of Economic Policy 21*(3), 416–438.

Fortin, N. M. (2015). Gender role attitudes and women's labor market participation: Opting-out, aids, and the persistent appeal of housewifery. *Annals of Economics and Statistics* (117-118), 379–401.

Foster, G. and L. S. Stratton (2018a). Do significant labor market events change who does the chores? Paid work, housework, and power in mixed-gender australian households. *Journal of Population Economics 31*(2), 483–519.

Foster, G. and L. S. Stratton (2018b). Does female breadwinning make partnerships less healthy or less stable? Technical report, IZA.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *The American Economic Review 104*(4), 1091–1119.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2018). Bartik instruments: What, when, why, and how. Technical report, National Bureau of Economic Research.

Görges, L. (2018). Wage earners, homemakers, and gender identity–Using an artefactual experiment to understand couples' labour division choices. Technical report.

Gottesdiener, L. (2014, October). I worked in a strip club in a North Dakota fracking boomtown. *Mother Jones*.

Griliches, Z. and J. A. Hausman (1986). Errors in variables in panel data. *Journal of Econometrics 31*(1), 93–118.

Hederos Eriksson, K. and A. Stenberg (2015). Gender identity and relative income within households: Evidence from sweden. Technical report.

Hirsch, B. T. (2005). Why do part-time workers earn less? The role of worker and job skills. *ILR Review 58*(4), 525–551.

Manning, A. and B. Petrongolo (2008). The part-time pay penalty for women in Britain. *The Economic Journal 118*(526), F28–F51.

McGinnity, F. and P. McManus (2007). Paying the price for reconciling work and family life: Comparing the wage penalty for women's part-time work in Britain, Germany and the United States. *Journal of Comparative Policy Analysis 9*(2), 115–134.

Murray-Close, M. and M. L. Heggeness (2018). Manning up and womaning down: How husbands and wives report their earnings when she earns more. *US Census Bureau Social, Economic, and Housing Statistics Division Working Paper* (2018-20).

Page, M., J. Schaller, and D. Simon (2019). The effects of aggregate and gender-specific labor demand shocks on child health. *Journal of Human Resources 54*(1), 37–78.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2018). Ipums usa: Version 8.0 [dataset]. *Minneapolis, MN: IPUMS*.

Sanderson, E. and F. Windmeijer (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics 190*(2), 212–221.

Zinovyeva, N. and M. Tverdostup (2018). Gender identity, co-working spouses and relative income within households. Technical report, IZA.

A BKP Replication

In this section, I replicate BKP's results and compare with modified specifications. In the modifications, I adapt the key measure of the probability that the wife earns more than the husband based on the insight that this measure may be biased for women who actually earn more than their husbands. Using these adapted measures, the association between the probability that the wife earns more and labor force participation is either positive or insignificantly different from zero.

According to BKP's notion of the breadwinner norm, we may expect that couples reduce a wife's labor supply or participation if she is likely to outearn her husband on an annual basis. By doing so, they would avoid violating the norm.

To test this relationship, BKP construct a measure of the probability that a wife would outearn her husband as follows.

First, they assign every woman (regardless of working status) a distribution of potential earnings by calculating the vigintiles of the annual earnings distribution for the working women with the same state, age group (five-year intervals), race, and educational attainment (four levels), and year. Denote each vigintile of earnings as $\hat{y}_w^i$, for $i = 1, ..., 19$.

Next, they compare each of the moments of the assigned distribution for a given wife to her husband's *observed annual earnings*, generating a value of one if a given moment for the wife exceeds the husband's earnings. Finally, they average across the 19 moments. To be concrete, this measure is computed as:

$$\text{PrWifeMore} = \frac{1}{19}\sum_{i=1}^{19} \mathbb{1}(\hat{y}_w^i > y_m)$$

A key feature of this measure is that it incorporates the hours decision as part of the wife's potential earnings. Depending upon the distribution of women working part-time or full-time in a given demographic group, individual estimated earnings (and thereby the probability that the wife earns more) might be upwards or downwards biased. *A priori*, it is not obvious which direction this bias might have on average. However, there is one group in particular for whom this bias will be biased downwards: women who earn more than their husbands.

First, I will note that the sample I use is different with respect to years: BKP include decennial Census waves from 1970 to 2000, and then use a pooled sample from 2008-2010 as their most recent year, while my sample does not include 1970 and has a pooled sample of 2005-2007 as the most recent wave.

If we observe a wife actually earning more than her husband, it seems reasonable suppose that the probability of earning more than her husband should be high. In fact, the average (and median) value of *PrWifeMore* is 0.36 for the couples where the wife is actually observed to earn more. While this is higher than the average value of *PrWifeMore* for couples where the wife in fact earns less or does not work, it nevertheless seems likely that it might be biased downwards relative to the true probability that they outearn their husbands. Naturally, this fact depends somewhat on what we mean by the "true probability". Therefore, I consider a few different ways of redefining *PrWifeMore*.

In the replication, I denote the original variable by λ_1. One way to modify their key variable would then be to recode this measure with a value of one for women who actually earn more than their husbands (λ_2), given that we observe with certainty that they do outearn their husbands. However, it's somewhat inconsistent on the one hand to treat men's earnings and one subset of working women as having certain outcomes, while treating women who earn less than their husbands differently. Another variation I consider is treating all of those individuals who have observed income as having certain outcomes: Thus, women who earn less than their husbands are coded as having zero probability of earning more, while women who earn more have a probability of one, and those who do not work retain the original value (λ_3). Finally, one might reasonably think that even if a given outcome is observed in this time period, there is uncertainty in subsequent time periods. However, this point applies even to men's earnings. In a third and final variation, I estimate a distribution for men's wages just as described for women, and construct the probability the the wife earns more by comparing each moment of their respective distributions and averaging across the full distribution (λ_4).

$$\lambda_4 = \frac{1}{19} \sum_{i=1}^{19} \mathbb{1}(\hat{y}_w^i > \hat{y}_m^i)$$

I will note that under this measure, women who are currently observed to earn more than their husbands are assigned even lower probabilities of doing so: the average value of λ_4 for wives who earn more is 0.20, and the median is 0.17.

I follow the same specification as BKP use in Col. 3 of Table 2 in their paper.[13] This specification includes fixed effects for husband and wife's educational attainment level, age group, state, and race, year fixed effects, a dummy for whether the couple has children, the full set of vigintiles for the wife's predicted earnings, and (for λ_1 to λ_3) a cubic in the husband's log income and an interaction between the husband's log income and the median predicted earnings for the wife. These latter controls that involve the husband's observed income are excluded when the probability measure also involves predicted earnings for the husband. Instead, I include the full set of vigintiles for the husband's income, as is done for the wife. Standard errors are clustered at the wife's demographic group level, except for λ_4 where the group level is defined as the combination of the husband and wife's demographic

[13] I don't use the specification in Col. 4 because husband and wife group fixed effects together fully capture the variation in λ_4. For the other three versions, I have tested this additional specification and the results are very similar.

Table A.1: BKP Replication and Variations on $ProbWifeEarnMore$

(a) Labor force participation

Dependent Variable: Wife's LFP	(1)	(2)	(3)	(4)
λ_1	-0.164*			
	(0.006)			
λ_2		0.246*		
		(0.003)		
λ_3			0.059*	
			(0.002)	
λ_4				-0.004
				(0.011)
Observations	4,487,547	4,487,547	4,487,547	4,487,547
R-squared	0.089	0.108	0.090	0.072

Robust standard errors in parentheses
* p<0.01

(b) Hours

Dependent Variable: Log(Weekly Hours)	(1)	(2)	(3)	(4)
λ_1	-0.065*			
	(0.006)			
λ_2		0.245*		
		(0.003)		
λ_3			0.223*	
			(0.002)	
λ_4				-0.115*
				(0.012)
Observations	3,410,714	3,410,714	3,410,714	3,410,714
R-squared	0.051	0.076	0.082	0.046

Robust standard errors in parentheses
* p<0.01

group.

I provide results for both labor force participation and hours as the dependent variable, where hours is the log of usual weekly hours worked, as given in BKP's online appendix.

Beginning with labor force participation, the main focus in BKP, Col. 1 of Panel (a) in Table A.1 gives the result of exactly replicating BKP, given my sample. The coefficient on λ_1 is -0.16, which is similar but slightly more negative than what they find (about -0.14). Similarly, in Panel (b) the coefficient is -0.069, also more negative than the value of -0.034 which they estimate (in Online Appendix Table A.8). The results in both cases would yield the same conclusions despite differences in the exact point estimates.

In Col. 2, where the couples with higher earning wives have $\lambda_2 = 1$ and all others have $\lambda_2 = \lambda_1$, the coefficient is still significant and turns positive, with a value of 0.249 on labor force participation and 0.247 on the log of weekly hours. In other words, it implies a 10 percentage point increase in the probability that the wife earns more by this measure is associated with a 2.49 percentage point increase in the probability of working at all, and a 2.47 percentage point increase in weekly hours, conditional on working.

In Col. 3, where $\lambda_3 = 0$ for all working wives who earn less than their husbands, and $\lambda_3 = \lambda_2$ for all others, the coefficients are still positive. For labor force participation, it is smaller in magnitude, at 0.058, while for hours it is relatively similar as for λ_2, at 0.22. Finally, for Col. 4, which estimates the husband's potential earnings in the same way as for the wife, the coefficient on labor force participation is a precisely estimated zero (for an increase of 10 percentage points in λ_4, the 95% confidence interval is between 0.00028 and 0.00043 percentage points). However, the coefficient on the log of weekly hours is negative and larger in magnitude that for the original measure λ_1, at -0.115.

Given that the alternative measures yield coefficients that are different in sign, not just magnitude, I do not believe sample differences are driving these results.

These results suggest that BKP's original findings are quite sensitive to bias in the key measure. The fact that the sign and magnitude vary so greatly for logical variations on the key variable suggests that this relationship should be reexamined.

B Census Rounding Procedures

Fig. 1 replicates the distribution of relative income in the U.S., as produced in BKP. As BKP note, there are disproportionately many couples that appear to have exactly equal annual incomes in the Census data when compared with administrative tax data for couples that both have positive income: roughly 3% for each year exhibit this feature in the Census data, whereas only about a quarter of a percent do so in the administrative data (Survey of Income and Program Participation Gold Standard Files). Couples where either member are self-employed, which might lead to simply splitting total family income across the two individuals, are already excluded in my sample. Moreover, couples that both have top-coded income are dropped, but there are fewer than 100 such cases.

Thus, BKP determine that much of the remaining excess couples with identical incomes may be due to rounding issues. I initially follow BKP in derounding the annual earnings variable for men and women respectively according to IPUMS' description of comparability across years for this variable (Ruggles et al., 2018), which is as follows:

"For 1980, the codes represent the midpoints of ten-dollar intervals. The 1940 and 1990 codes are expressed in exact dollar amounts instead of intervals.

The 2000 census, the ACS and the PRCS codes are amounts rounded as follows: No income \$0 \$1 - \$7 \$4 \$8 - \$999 rounded to nearest \$10 \$1,000 - \$49,999 rounded to nearest \$100 \$50,000 or more rounded to nearest \$1000"

These rules appear to be accurate for 1980 and 2000: Couples with equal earnings are virtually eliminated through derounding by adding a draw from the uniform interval that could have given rise to the rounded value (0.01% of such couples remain for 1980 and precisely zero for 2000). However, more than 3% of couples in 1990 still have equal earnings. The contrast suggests that 1990 values may also be rounded, and direct inspection of a number of observations supports this idea.

Since I do not have rounding rules to follow for derounding for 1990 values, I apply a similar derounding procedure if a given value is divisible exclusively by 1,000, 100, or 5 respectively.

Following this adjustment, only 0.01% of couples in 1990 have equal earnings, consistent with the results for the other years.

C Assessing the Validity of the Exclusion Restriction

For the Bartik IV strategy to be valid in estimating the causal effect of the relative wage on wives' hours, the exposure shares, defined at the industry-educational attainment level, should not be related to wives' hours except via their effect on the relative wage. A concern might be that industry shares could be correlated with gender norms, which may impact both the relative wage and hours via a relationship to selection into marriage.

To provide some evidence on whether there is reason to be concerned that industry shares are correlated with gender norms, I use data from the General Social Survey on responses of agreement or disagreement to the statement, "It is much better if the man is the achiever outside the home and the woman takes care of the home and family." While this does not necessarily capture the "breadwinner norm" directly in the sense of the concern with relative earnings, it does speak to closely related views on gender roles. Like in the Census, this dataset also records the industry in which respondents are working. Thus, I can test for differences in the agreement with the norm across industries for individuals of the same educational level and sex. If indeed the industry shares are correlated with views on norms, it must be that individuals working in some industries are more traditional with respect to norms than individuals in others. However, it's possible that such an effect would owe to the average characteristics of individuals in a given industry, such as the share of a given educational level. Thus, a simple analysis of the views of individuals by industry and education can be informative on this point.

While the availability of direct responses to questions about norms are a key advantage of the GSS, a drawback its relatively small sample size compared to the Census, with about 60,000 observations over all years between 1978 and 2015. Respondents report industry in detailed categories consistent with Census classifications, but there are too few responses to usefully compare across the industry-education categories as I define them in the Census (of which there are 972). Thus, I aggregate industries in 13 groups (as categorized by IPUMS; I exclude the military).

Responses to the statement are coded on a scale from 1 to 4 with higher values indicating agreement,[14] which I recode to a simple dummy measure of agreement or disagreement.

I estimate regressions of the following form:

$$y_{it} = \beta_0 + \beta_1(Industry_{it}) + \beta_t(Year_t) + \epsilon_{it}$$

Since this survey has been conducted over many years, this particular question has responses from 1978 to 2015. To maximize sample size and given that all groups have exhibited marked declines in agreement with this norm over time, I pool responses across years under the assumption that the full set of year dummies will capture the common time trend.

I estimate this regression separately by sex and for each industry in turn, to obtain the coefficient β_1, which captures the extent to which individuals working in an industry i are more or less in agreement with individuals in all other industries with respect to the norm.

Fig. C.1 graphs the coefficients on industries by sex with 99% confidence intervals, to account for the multiple hypotheses being tested. Several industries do indeed appear to be significantly more or less in agreement with the norm compared to others: in particular, men in agriculture, construction, and transportation are estimated to be more traditional (more in agreement with the norm) than others while men in arts and entertainment, professional services, education and health, and public administration are less traditional than others. Meanwhile, women in agriculture, "Other services", and manufacturing are likewise more traditional than other women, and those in professional services and educational and health are less traditional.

[14]The responses correspond to the following choices: "Strongly disagree" (1); "Disagree" (2); "Agree" (3); "Strongly Agree" (4).

Next, I add to the regression a dummy variable for whether the individual has at least some college education or more, and interact this dummy with the individual's industry:

$$y_{it} = \beta_0 + \beta_1(Industry_{it}) + \beta_2(Edu_{it}) + \beta_3(IndXEdu_{it}) + \beta_t(Year_t) + \epsilon_{it}$$

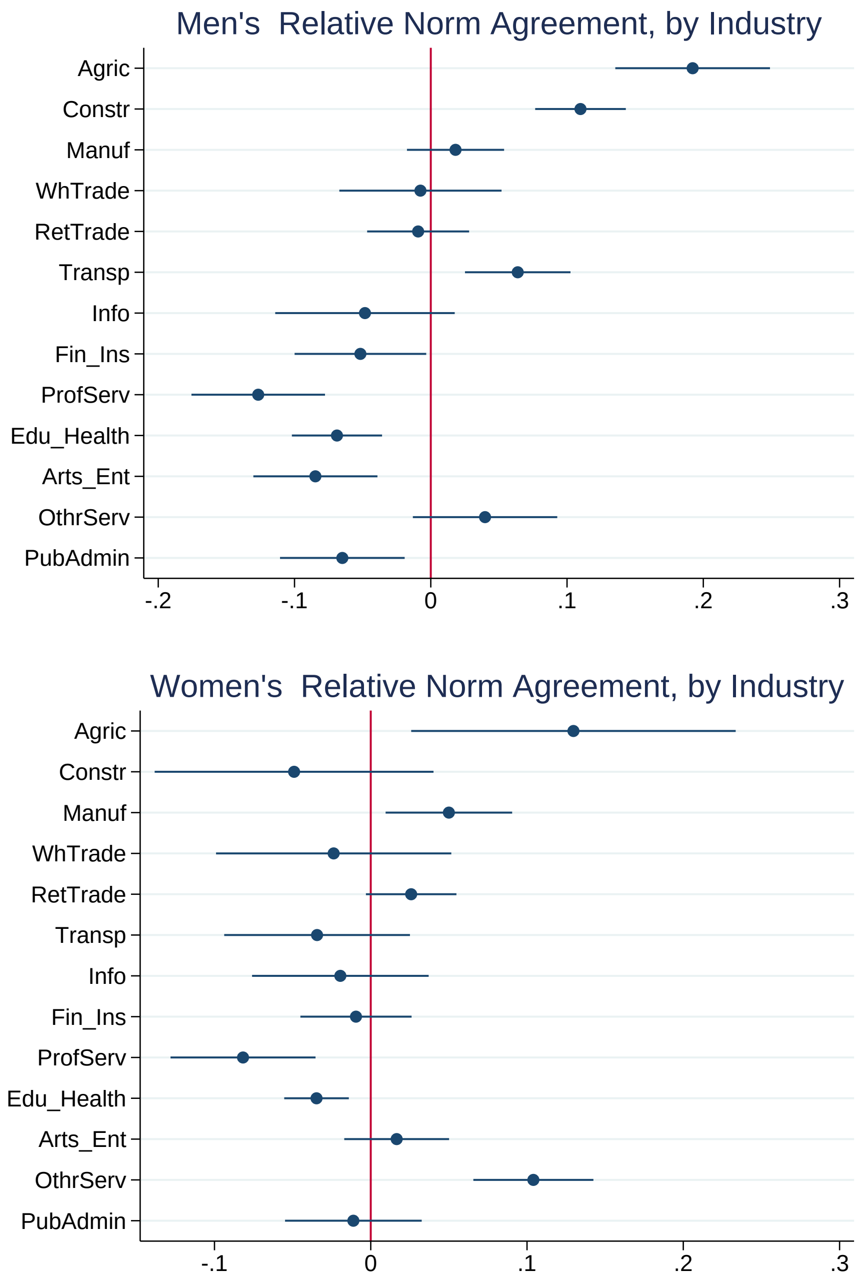
Here, the omitted group is individuals in other industries with a high school degree or less. Thus, β_1 captures whether the lower-educated individuals in industry i differ with respect to the norm from the lower-educated in other industries, β_2 indicates how the higher-educated individuals in all other industries differ from the lower-educated, and β_3 gives the extent to which the higher-educated individuals in industry i differ in their agreement from those in the other industries.

Fig. C.2 graphs the coefficients β_1 and β_3 by industry. In contrast to the aggregate differences across industries, there are fewer significant differences within educational groups. However, men in agriculture with a high school degree or less are still significantly more traditional than their peers. Moreover, higher-educated men working in construction and manufacturing also appear to be more traditional, while lower-educated men in arts and entertainment are still less traditional than others.

For women, the picture is more straightforward: only lower-educated women in "other services" appear to be slightly more traditional than similarly-educated women. The higher-educated women have no notable differences.

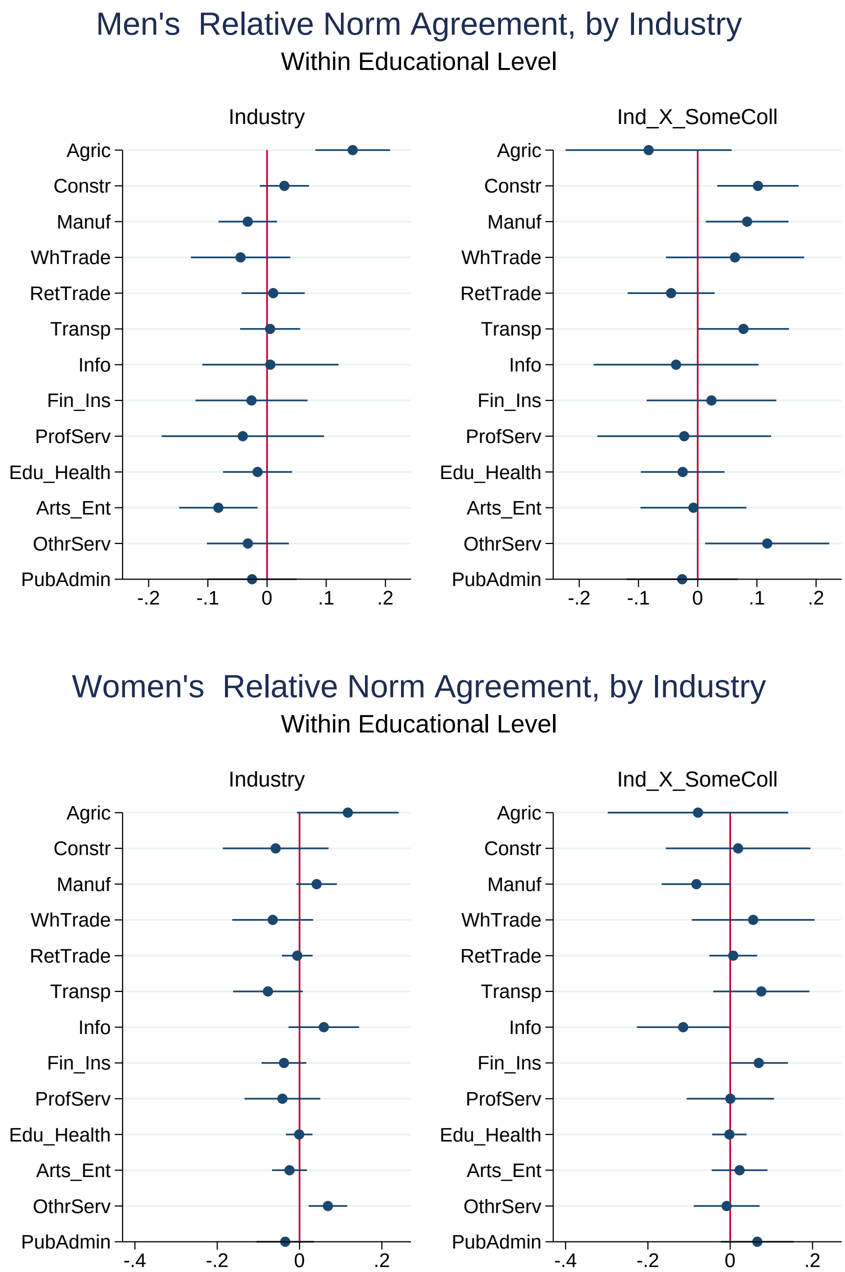
This analysis provides some reassurance that there are relatively few differences across industries in gender norms conditional on educational attainment. However, to the extent that some differences are still evident, when I analyze the identifying variation in the Bartik instruments using Rotemberg weights, these results provide some clues as to which industries might be a cause for concern if they account for large shares of the variation.

Figure C.1: Agreement with Gender Norms by Industry and Sex



Using data from the GSS, each marker plotted with 99% confidence intervals represents a coefficient from a single regression of agreement with the gender norm on the indicated industry dummy and a full set of year dummies from 1978 to 2015.

Figure C.2: Agreement with Gender Norms by Industry, Sex, and Education



Using data from the GSS, each marker plotted with 99% confidence intervals represents a coefficient from a single regression of agreement with the gender norm on the indicated industry dummy (or the industry dummy interacted with the education dummy) and a full set of year dummies from 1978 to 2015.

D Bartik Instruments with a Quadratic First Stage

When considering the specification with the relative wage and the total household wage as endogenous regressors, the Bartik shocks for men and women respectively could in principle be sufficient for just-identified estimates with 2SLS.

When estimating the first stage as such, using linear Bartik shocks as instruments, both the standard F-stats and AP F-stats imply rejection of the null of that each endogenous regressor is weakly identified. However, Sanderson and Windmeijer (2016) show that both standard F-stats and even the modified Angrist-Pischke F-stats are likely to overreject as tests of weak identification in a setting with multiple endogenous regressors where the matrix of parameters of the reduced-form equations has a near-rank reduction of one. They provide new conditional F-statistics that correct for these issues, hereafter refered to as SW F-stats. The SW F-stats for the linear specifcation with linear Bartik shocks imply that the null of weak identification for each endogenous regressor is not rejected (results available upon request)

It's not entirely surprising that this issue arises, as the two endogenous regressors are closely related, though not linearly dependent, and hence their relationships to the instruments may not be sufficiently different to separately identify the coefficients on each regressor.

The squares of the Bartik shocks are valid instruments if the Bartik shocks themselves are mean independent of the error term. While this assumption is stronger than uncorrelatedness needed for consistency using only the linear Bartik shocks, as (Dieterle and Snell, 2016) point out, it's difficult to think of an economic setting in which we would have uncorrelatedness and not mean independence. Moreover, they show that including the square of the instrument and testing the resulting overidentification restriction can in fact be a diagnostic for whether an instrument is valid. Under the assumption of mean independence between the instrument and the error term, a rejection of the overidentification test when including the square of the instrument in the first stage would imply a failure of the mean independence assumption: if the linear Bartik shock is a valid instrument that is mean independent, its square should be too. A rejection implies that one of the instruments is not valid.

Using the quadratic first stage, Table **??** shows the Hansen J statistic and associated p-value for the test of overidentification of the first-stage instruments when adding the squares of the Bartik shocks as instruments. They overidentification tests are not rejected either for the relative wage or the total household wage. As the SW F-stats also indicate rejection of the null of weak identification for each endogenous regressor, these tests provide some reassurance that using both the linear and quadratic Bartik shocks as instruments is a valid strategy.

E Inference

In applications using geographical units, it is typical to cluster standard errors by these same geographical units. The assumption is that only areas that are in physical proximity have correlated errors, while those that are geographically distant are necessarily independent. However, Adão et al. (2018) show that in the case of shift-share instruments, such standard errors may yield inappropriately small confidence intervals. The potential issue they highlight is that even geographically distant areas may have similar distributions of industries, and hence have correlated (potential) outcomes. While the structure of my analysis differs such that I cannot directly apply their proposed correction, I can test the extent to which this issue is present by conducting the same placebo test they use to motivate the paper.

The test involves constructing placebo Bartik instruments by replacing the observed industry wage shocks by randomly generated shocks drawn from a normal distribution with zero mean and the same standard deviation as the observed distribution of wage shocks across all industries (0.12). Then, I estimate the 2SLS equations exactly as in Table 8, Col. 3. I repeat this procedure 2,500 times.

The results of the test show that there may be some degree of overrejection to due correlation across geographical units, but the issue is much smaller than in the original application that Adão et al. (2018) discuss. In this setting I find a rejection rate of 8.3% for the 5% significance level, 1.5% at the 1% level, and 0.16% for the 0.1% level. In comparison, Adao, Kolesar and Morales reject at a rate of about 45% for a 5% significance level in the paper they replicate. Thus, while chance rejection of the null does occur at slightly higher rates than would be typically expected for these hypothesis tests, it is not nearly as severe of an issue as might be the case with other shift-share applications.