Earnings Inequality and the Changing Association between Spouses' Earnings in Urban China (1988-2002)

Yifan Shen

The Chinese University of Hong Kong

Full paper—PAA 2015

Abstract

China witnessed dramatic increase in earnings inequality in the past several decades. Little is known, however, about the role of marital sorting in shaping this rise. How much of the observed increase in earnings inequality among married couples in urban China could be attributed to changes in the association between spouses' earnings? To answer this question, this paper uses data from CHIP (Chinese Household Income Project) 1988, 1995 and 2002 and four different methods in existing literature. I find that, after evaluating the deficiencies in each method and clarifying the composition of the association structure, I find that the impact of changes in the association between spouses' earnings on trends in inequality in urban China is negligible. Only one component of the association——changes in the pattern of wives' labor force participation——have moderate disequalizing impact.

Introduction

Income inequality surged in the past three decades in China (Li et al. 2013). This research aims to understand the mechanisms generating the increases in income inequality by focusing on the role of marriage and family patterns.

It will contribute to two sets of existing literature. First, it will provide empirical proof on an assumption that is still only assumed to be true in current literature on assortative mating in China. One rationale of analyzing trends in marital sorting on socioeconomic traits lies in its potential consequences on the correlation between spouses' income which, in turn, will arguably affect income inequality across families. It is found that there have been significant increases in the

trends of educational and occupational homogamy in China during the reform era, which is interpreted as a sign of deterioration in societal openness, but we do not know whether this interpretation or inference is proper. This study will shed light on the plausibility of one of the many paths through which homogamy influences societal openness by quantifying the impact of changes in the association between spouses' earnings on the rise in family income inequality.

Second, this research provides new insights into the role of changes in family patterns in the generation of income inequality. It offers a more precise measurement of the contribution of the association between spouses' earnings. Previous studies have found that family structure (whether a family is composed of a couple/single person with/without dependent members) has negligible impact on the magnitude of inequality at a point in time or on the trends in inequality in (urban) China (Wu 2010; Xie & Zhou 2014). However, another family pattern---the association between spouses' earnings or income---has received little research yet. After all, inequality across families depends on not only the dispersion of income within men and women but also how men and women are sorted into families. Ding et al. (2009) is the only work to my knowledge that has measured the contribution of changes in spouses' earnings correlation on the rise in family income inequality in urban China. They found that increases in the correlation between wives' earnings and other family incomes contributed to 6.3% of the total increase in family income inequality in urban China from 1988 to 1995 and 0.3% between 1995 and 2002. My study improves their work by estimating the impact of changes in earnings' correlation between wives and husbands instead of 'other family incomes' that include not only husbands' earnings but all residual family incomes.

More importantly, because the size (and possibly even direction) of the effect has been found to be dependent on the methods and measures used by different researchers, I shall make use of multiple methods, compare their pros and cons and see if there is any consistent finding.

Research Question

I do not intend to make causal inference. Instead, I attempt to estimate the impact of changes in the association between spouses' earnings on trends in family earnings inequality across a period of time, which is virtually a bivariate correlation. I consider the counterfactual that how much earnings inequality among married couples would have changed if there were no change in the association between spouses' earnings across the period? I will use four different methods to answer this question to see how the results resemble or differ from each other and why. Where the method permits, I will take a closer look at the question. First, I decompose the association between spouses' earnings into different dimensions and components and estimate their separate impact on trends in earnings inequality. Second, I use multiple measures of inequality to see not only how trends in overall inequality were affected but also how the impact differs at different parts of the earnings distribution.

In this draft I will concentrate on the role of changes in the association between spouses' earnings (instead of income) in explaining trends in inequality of couples' earnings (instead of family income) in urban China (instead of the whole country). Previous literature found that change in the female labour force participation plays a significant role in shaping trends in income inequality in urban China. Because the annual earnings of the unemployed are naturally zero, focusing on laboured earnings instead of income allows me to examine the independent role of female labour force participation. In total, the share of the earnings of couples in total family income (family where at least one couple is present) is 84% in 1988, 85% in 1995 and 74% in 2002. In the final section of this draft I will discuss how other issues can be addressed within my framework, i.e. family size adjustment, how to examine the impact of changes in the earnings association on trends in family income as opposed to the earnings of couples only, how to incorporate single-person households, etc.

Data

I use household income survey data from CHIP 1988/1995/2002. Below is some basic information of the three samples used.

Table 1 Data Summary

	Number of Observations (couple)	**Of which: Dual earners**
1988	7013	6376
1995	5217	4405
2002	5200	3260

Table 2 Descriptive Statistics of Earnings of Married Couples (unweighted)

	1988		1995		2002	
	proportion	mean	proportion	mean	proportion	mean
Total	1.00	10069	1.00	13275	1.00	18912
East	39.75	11372	35.79	16635	36.15	23730
Middle	43.16	8757	35.71	11016	36.37	15351
West	17.08	10349	28.50	11886	27.48	17286
N		7013		5217		5200

Note: all mean earnings in the table refers to 2002 Yuan

Only households where 1) both partners are present; 2) wives are aged between 16 and 55 and 3) husbands are aged between 16 and 60 are included. When more than one couple is present within one household, only the one that contains the household head is included. Earnings of husbands and wives are defined as 'the sum of wages (salaries), non-wage compensation, net income of self-employment and non-monetary benefits minus income tax' (Ding et al. 2009). All income measures are discounted by urban CPI (obtained from the online published data of NBS website) with 2002 as base year.

Methods

One (M1): Decomposing CV by source (Cancian & Reed 1999)

The squared coefficient of variance (CV2) can be decomposed as below:

$$CV_c^2 = a^2 CV_h^2 + b^2 CV_w^2 + 2\rho ab CV_h CV_w \quad (1)$$

where CV_c, CV_h and CV_w are the coefficient of variation of earnings of couples, husbands and wives, respectively; a is the share of husbands' earnings in couples' total earnings, b the share of wives'; ρ is the coefficient of correlation between spouses' earnings. It is thus straightforward that a higher ρ will amplify couples' earnings inequality if all else is held constant. Therefore, the CV_c^2 under the counterfactual that only the association between spouses' earnings had not changed during a period (1988-2002 here) can be easily worked out by setting ρ to its 1988 value while keeping all other factors (CV_h, CV_w, a and b) at their 2002 levels. This method has been widely used in existing literature.

Two (M2): Non-parametric approach (Burtless 1999; Reed & Cancian 2012)

Unlike M1, this method focuses on the *rank correlation* between spouses' earnings instead of the coefficient of correlation. In the first step, the annual earnings of husbands and the annual earnings of wives are divided into 1000 equal-sized categories in 1988 and 2002. Second, two new distributions are simulated to represent the observed distributions of husbands' and wives' earnings in 2002. In the simulated distributions, the earnings of a husband (wife) in millicile i equal the mean of the earnings of all the husbands (wives) in the same millicile. In doing so, we assume that 1000 milliciles suffice to represent the inequality of the original distribution, which is empirically confirmed (Gini, CV, p90/50, p50/10 are all extremely similar). Third, we replace the earnings of each husband in millicile i (i=1,2,…,1000) in 1988 by the earnings of husbands in the

same earnings rank (millicile i) in the 2002 simulated distribution. In this way, we construct a counterfactual distribution in which the earnings dispersion of husbands and wives are at 2002 level but the whole rank correlation is kept as it was in 1988. The difference in inequality between the (simulated) observed 2002 distribution and the counterfactual distribution thus indicates the size of the impact of changes in the association between spouses' earnings. Because the counterfactual distribution is simulated, we are able to use multiple measures of inequality to explore different trends in various portions of the earnings distribution, which is a great advantage over the first approach which relies solely on CV.

Both the first two methods share one deficiency: the association between spouses' earnings is simplified into a single measure, no matter whether it is a single summary measure like the coefficient of correlation, or a rank correlation which can only be preserved as a single integrity. However, sometimes we are equally interested in the structure of association. There are at least two components of the association between spouses' earnings ranks: 1) the association between spouses' earnings ranks among dual-earner couples, and 2) the relationship between one partner's earnings rank and the likelihood that the other one does not work (has zero annual earnings). Such decomposition is important especially in the context of urban China where the female labour force participation plays an important role in the generation of family income inequality during the market transition. Two methods that enable such decomposition are introduced below.

Three (M3): Additive Decomposition (Breen & Salazar 2011)

This approach makes use of the additively decomposable property of the generalised entropy index. Because in this draft I focus on the annual earnings, a significant minority of couples in the data consist of partners both of whom are unemployed, i.e. have zero earnings. Therefore, I use GE(2) only. This also makes the result of this method comparable to the first two methods, because GE(2) is just half the squared coefficient of variation. Firstly, I divide the earnings of

husbands into six categories, with one category for those with zero earnings and the others for the five quintiles of those with non-zero earnings, i.e. the bottom group consists of husbands whose earnings are in the bottom quintile (1-20%), the second for those in 21-40%, and so on, and repeat the same procedure on wives earnings. A cross-tabulation of spouses' earnings categories results in a 6 X 6 table, which classifies all couples in each year into 36 types. In this way, the overall inequality of earnings among couples ($GE(2)^T$) in year i is then decomposed into two parts: inequality between the 36 couple types ($GE(2)^B$) and inequality within each couple type ($GE(2)^W$):

$$GE(2)^T = GE(2)^B + GE(2)^W$$

$$= \frac{1}{2\bar{y}^2}\sum_{k=1}^{36}(\bar{y}_k - \bar{y})^2 p_k + \sum_{k=1}^{36}(\frac{\bar{y}_k}{\bar{y}})^2 GE(2)_k\, p_k \quad (2)$$

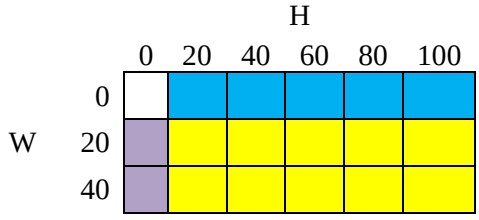
where $\bar{y}$ is mean earnings of all couples, $\bar{y}_k$ is mean earnings of couples in couple type k (k=1,2,…,36), $GE(2)_k$ is within-group $GE(2)$ for couple type k, and p_k is the proportion of couples in couple type k among all couples. Because the overall mean earnings $\bar{y}$ is actually the sum of mean earnings in each couple type weighted by the relative size of each couple type p_k, Equation (2) can also be written as below:

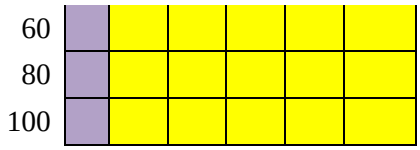
$$GE(2)^T = \frac{1}{2(\sum_{k=1}^{36}\bar{y}_k p_k)^2}\sum_{k=1}^{36}(\bar{y}_k - \sum_{k=1}^{36}\bar{y}_k p_k)^2 p_k + \sum_{k=1}^{36}(\frac{\bar{y}_k}{\sum_{k=1}^{36}\bar{y}_k p_k})^2 GE(2)_k\, p_k \quad (3)$$

from which we see that the overall level of inequality, when measured by GE(2), can be seen as a function of three quantities: the within-group $GE(2)_k$; the couple type mean earnings $\bar{y}_k$; and the distribution of couple types p_k. To assess the contribution of the association between spouses' earnings to overall inequality, I calculate $GE(2)^T$under the counterfactual that p_k is set to their 1988 values while the other two quantities remain at their 2002 levels.

Because p_k are actually cell proportions of the 6 X 6 contingency table, we are able to change/keep them in various ways so as to decompose changes in the association in the contingency table into different components. The association can be decomposed into three parts: the association between spouses' earnings categories among dual-earners (component 1, or C1 below), which corresponds to the yellow area in Figure 1; the relationship between husbands' earnings categories and the likelihood that wives work (component 2, or C2 below), which corresponds to the blue area; the relationship between wives' earnings categories and the likelihood that husbands' work (component 3, or C3 below), the purple area. The cell for dual-non-earners can be included in either the second or third component, and my analysis shows that whether it is included or not does not affect the results substantially. To quantify the contribution of changes in the association among dual-earner couples, I replace the relative frequencies of the yellow area of the table for 2002 couples using those from the same area on 1988 table, while keeping the other cells at their 2002 value. The marginal distribution of the new table is not necessarily equal to that of the 2002 table, nor is the sum of cell proportions equal to one. I use Deming-Stephen (1940) algorithm to adjust the cell proportions to keep the marginal distribution of the new table consistent with the observed 2002 table. The GE(2) calculated using this counterfactual table thus represents the overall level of inequality in 2002 if only the association between spouses' earnings categories among dual-earners had not changed from 1988. Similarly, to quantify the contribution of changes in the other two components, the blue or purple area is replaced instead.

Figure 1





In summary, this approach models changes in the association between spouses' earnings by comparing the distribution of cell proportions of the 6 X 6 tables between any two years. Because decomposing GE(2) entails the calculation of within-group GE(2), which is based on all the observations available within each cell, small cells have to be avoided. Therefore, when applying this approach the earnings of working partners are classified into only five quintiles.

Four (M4): log-linear models (Schwartz 2010)

Similar with the third method, the fourth approach cross-tabulates the earnings categories of husbands and wives into contingency tables. Specifically, I classify earnings into 11 categories, with the earnings of working partners being divided into 10 deciles and 1 particular category for those with zero earnings. I cross tabulate the husbands' and wives' earnings categories in 1988, 1995 and 2002 into an 11X11X3 contingency table. Then I use log-linear models to model changes in the association between survey years. Formally the baseline model can be written as below:

$$\log \mathrm{F}_{\mathrm{ijt}} = \lambda + \lambda_i^H + \lambda_j^W + \lambda_t^Y + \lambda_{it}^{HY} + \lambda_{jt}^{WY} + \lambda_{ij}^{HW} \qquad (4)$$

The cross-sectional association λ_{ij}^{HW} is saturated to improve the goodness-of-fit of the model so that I can focus on terms for the interaction between HW and the layer variable (i.e. years). Table 3 is the goodness-of-fit statistics of ten different model specifications I have tried. Model 1 is the baseline model as specified in Equation (4). In model 2 I use the mean earnings of all members within each earnings category as scales of the row and column of the contingency table and fit the

linear-by-linear association model. It turns out that such a linear-by-linear association does not improve the goodness-of-fit at all. In model 3 and 4 I try to see if changes in the distance between spouses' earnings categories (|Hp-Wp|) or in the distance between spouses' absolute earnings (|Hs-Ws|) help to explain variations in the association between dual-earners' earnings categories. Once again there is no improvement. Is it because the association between spouses' earnings categories changed little during the period? Table 4 supports this hypothesis. Models in this table are based on data of dual-earners only. In the first row, the 'homo rc1'---homogeneous RC model where only one dimension of RC scores is estimated---is fitted. Compared to the conditional independence model (not shown), the first dimension of homogeneous log multiplicative RC association explains 80% of the total deviance, while the first dimension of heterogeneous RC association explains 81% (shown in the second row 'het rc1'). Similarly, three dimensions of homogeneous RC association (RC scores vary across dimensions) explains 95% of the total deviance, and its heterogeneous counterpart explains only slightly more (97%). If we assume that three dimensions suffice to represent the whole association, only 2% of the association in the first three dimensions come from the heterogeneous component. I also fit the log-multiplicative layer effect model where the cross-sectional association λ_{ij}^{HW} is saturated (Xie 1992; model statistics not shown). Compared to the full interaction (FI) model (not shown), this model does not improve model fitness at all. With all the proof I conclude that there is no need to add any other term for changes in C1---the association among dual-earner couples. Back to Table 3, I fit model 5 and 6 to see if there is any significant changes in C2---relationship between husbands' earnings categories and the likelihood that wives do not work. Model 5 fit all the cells where wives do not work exactly. It explains 48% of the total deviance in the baseline model, which is the largest decrease in deviance so far. However, it is not parsimonious (BIC bigger than the baseline model). In model 6 I specify the relationship to be quadratic. It consumes 14 less df than model 5 but explains roughly the same amount of deviance. The BIC is also in favour of this model. In model

7 and 8 I tried to improve upon model 6 by adding terms for changes in C3---relationship between wives' earnings categories and the likelihood that husbands do not work. Again it turns out the quadratic specification does better. However, the added terms for changes in C3 do not improve the model fitness. BIC still favours model 6, and the likelihood-ratio test suggests that model 8 does not differ significantly from model 6 (p=.22). Therefore, I conclude that there is basically no substantial change in the degree or pattern of association in C1 and C3 from 1988 to 2002, but there is significant change in C2. Model 6 thus becomes my *preferred model*.

Table 3

	Model	Deviance	df	Delta	BIC	p	Ratio
1	Baseline	390.37	200	5.58	-1562.82	0.000	1.00
2	Baseline+L-by-L	389.99	198	5.56	-1543.67	0.000	1.00
3	Baseline+|Hp-Wp|	387.73	198	5.58	-1545.93	0.000	0.99
4	Baseline+|Hs-Ws|	389.16	198	5.57	-1544.50	0.000	1.00
5	Baseline+wzHp*y	242.22	180	3.90	-1515.65	0.001	0.62
6	Baseline+wzHp2y	252.85	194	4.22	-1641.74	0.003	0.65
7	model 6 + ZhWY	224.66	176	3.94	-1494.15	0.008	0.58
8	model 6 + ZhW2Y	247.10	190	4.18	-1608.43	0.003	0.63

Table 4 components of association

model used	Deviance	%
homo rc1	1177.44	0.80
het rc1	1117.10	0.81
homo rc2	483.45	0.92
het rc2	385.29	0.93
homo rc3	315.20	0.95
het rc3	184.42	0.97

Next, I use the predicted frequencies from my preferred model (model 6) to simulate two new distributions to represent the observed distributions of couples' earnings in 1988 and 2002 respectively. Like M2 (non-parametric method), in the simulated distributions, the earnings of a

husband (wife) in category i equal the mean of the earnings of all the husbands (wives) in the same category.

[footnote: In doing so, I assume that one 'zero' category and ten deciles suffice to represent the inequality of the original earnings distribution of each sex, and 11X11X3=363 cells suffice to capture important changes in the association between spouses' earnings ranks without losing much information. On the one hand, this classification scheme is more detailed than M3 (additive decomposition) so that we can compare between them to see if 11X11 may capture any important information that is lost by 6X6 in M3. On the other hand, because people in the same category are forced to have the same amount of earnings in the simulated distributions, this method lost information on within-group inequality, which is retained by M3 since M3 does not simulate distributions. I do not use more detailed classification scheme to minimize potential loss of information due to limited sample size (if the earnings are classified into 21 categories instead of only 11, i.e. 0, 1-5%, 6-10%...96-100%, the proportion of small cells (<=5 observations) is 17% for 1988, 27% for 1995 and 38% for 2002).]

Based on the simulated distributions, I obtain a size of change in earnings inequality among couples between 1988 and 2002. Because this result comes from my preferred model, it should be almost the same as the observed size of change using grouped data, though there might be some difference between trends (size of change) in inequality calculated from grouped data and the original individual data, which is shown in Table 5. I focus on the last two columns on the right because only these statistics will be used in the following estimation. It appears that as far as the size of change is concerned, the Gini coefficient is the most robust measure, followed by CV. P90/50 and p50/10 are percentile ratios that measure high-middle and middle-low inequality, respectively. The row titled 'high-middle' refers to the share ratio that measure high-middle inequality: the ratio of the sum of earnings of couples in the top 20% to the sum of earnings of couples in the middle 60%. Similarly, the last row 'middle-low' refers to the share ratio that

measure middle-low inequality: the ratio of the sum of earnings of couples in the middle 60% to the sum of earnings of couples in the bottom 20%. Theoretically I would prefer share ratios to percentile ratios in the context of this study because the decomposition of trends in percentile ratios is sensitive to the categorization of earnings used in the log-linear models while the decomposition in share ratios is not. However, it seems that data grouping downplays the size of increase in the middle-low share ratio by 20%, but overplays the same increase in the middle-low percentile ratio (p50/10) by 20%. It downplays the increase in the high-middle share ratio by 11% and the p90/50 by 15%. It is hardly quantifiable how these inconsistencies due to data grouping are going to affect the following estimation.

Table 5 Model 6 predicted (grouped data) vs observed (individual data)

Index	1988 observed	1988 predicted	2002 observed	2002 predicted	*change observed*	*change predicted*
Gini	0.208	0.203	0.384	0.379	*0.176*	*0.176*
CV	0.454	0.380	0.768	0.701	*0.314*	*0.321*
p90/50	1.494	1.667	2.172	2.242	*0.678*	*0.575*
p50/10	1.577	1.543	3.652	4.049	*2.075*	*2.505*
high-middle	0.564	0.599	0.796	0.805	*0.232*	*0.206*
middle-low	5.163	4.804	13.578	11.491	*8.415*	*6.687*

Now I have the observed sizes of change in inequality (measured by multiple indexes) among couples predicted from my preferred model (model 6 in Table 3). Next, I predict another set of sizes of change in the same way as above but using predicted frequencies from the baseline model, where there is no term capturing changes in the association between spouses' earnings categories. This set of change sizes thereby represents the trends in inequality under the counterfactual that there had been no change in the association between spouses' earnings categories across this period. By comparing these two sets of sizes of change, I am able to estimate how much of the increase in earnings inequality (not only overall inequality but also high-middle and middle-low inequality) can be attributed to changes in the association between spouses' earnings categories.

In this study, because the data shows no significant change in the other two components of the association between earnings categories, the contribution of changes in the whole association between earnings categories is thus equal to the contribution of changes in C2.

Finally, when using predicted frequencies from the baseline model, we will still get some trends in the association between spouses' earnings, if measured by the coefficient of correlation. Schwartz (2010) argues that these 'residual' trends in correlation are due to changes in the share of dual-earner couples. The coefficient of correlation would become zero if the cross-sectional association term λ_{ij}^{HW} is removed from the baseline model (let us call this model the independence model here). Schwartz interprets the difference between the size of change predicted from the baseline model and the size of change predicted from the independence model as the result of the 'composition effects' (Simkus 1984), which in the American context is interpreted as the contribution of changes in the share of dual-earner couples to trends in inequality. This makes sense since in the US the share of dual-earners has been increasing since 1960s, and such increases are very likely to result in a closer association between spouses' earnings.

However, such an interpretation may not hold in the context of urban China. The 'composition effects' is only part of the story. The other part is changes in the marginal distributions, i.e. changes in the relative distances between earnings categories of husbands and wives. Furthermore, the composition effects are not necessarily equal to changes in the share of dual-earners. It may refer to other patterns of the composition of the joint distribution of earnings categories. In summary, the residual trends in correlation are actually the result of changes in the composition of the contingency table (whether more or less people are migrating from areas where the correlation is higher to areas where the correlation is lower) weighted by changes in the marginal distributions. Although it is generally more desirable to make a distinction between the two in the model and to remove the effect of marginal changes, it is infeasible here due to the model

limitations, before I can come up with any methodological innovations. So far we have four components of changes in the association between spouses' earnings: apart from the aforementioned three components C1, C2 and C3, we now have C4—trends in the residual correlation due to the composition effects weighted by marginal changes.

In summary, I use four methods in this study because each of them has advantages that the others do not have. To estimate the impact of changes in total association between spouses' earnings (that include the impact of trends in residual correlation), I use M1. To estimate the separate impact of changes in each component of the earnings association, I use M3 and M4. However, it is not clear how its classification scheme (11X11) affects the estimation because it relies on the grouped data and thus ignores inequality within earnings groups. There are two solutions: either increasing the complexity of the classification scheme to minimize the information lost when grouping data, or incorporating within-group inequality into estimation. The former leads to M2 which classifies earnings into 1000 milliciles, while the latter leads to M3 which uses an even rougher classification scheme (6X6) but is able to take into consideration within-group inequality. Because M3 and M4 use different classification schemes and not both of them take into consideration within-group inequality, their results are not necessarily the same.

Results

The results are presented in four steps. First, results on the impact of TOTAL association, i.e. how much the level of inequality would have changed had there been no change in the total association between spouses' earnings, including changes in all the four components of the association? Second, results on the separate impact of changes in the main (C1, C2 and C3) as well as the residual part (C4) of the association. Third, results on the separate impact of changes in each of the three components of the main aspect of the association (C1, C2 and C3).

Table 6 shows the impact of changes in total association estimated from the first and the forth method (M1 and M4). Results of observed trends of M4 are based on the preferred model (model 6 in Table 3) as opposed to the original grouped data because the preferred model fits the data so well that the difference is negligible. As mentioned above, the observed inequality calculated by M4 differs from that calculated from the original individual data because M4 uses grouped data. Table 6 shows that the M4 grouped data tends to underestimate the values of CV for both years. Nevertheless, the size of observed changes (Chg 1), which is key to our estimation, seems to be much less affected by which type of data I use. In total, changes in total association are estimated to account for 7-14% of the total increases in overall level of earnings inequality among couples in urban China from 1988 to 2002, depending on the method and measure used. In other words, the increase in the coefficient of variation (Gini coefficient) of the overall inequality from 1988 to 2002 would be 7% or 14% (12%) less than it was observed were there no change in the total association between spouses' earnings during this period. M4 also estimates the impact of changes in the total association on different portions of the earnings distribution. When measured by share ratios, 22% of the observed increase in high-middle inequality and 34% of the observed increase in middle-low inequality can be attributed to changes in total association. When measured by percentage ratios, the effect of total association on middle-low inequality is still disequalising, though in a larger size (57%). The effect on high-middle inequality, however, is estimated to be slightly equalizing: without changes in total association, changes in the value of p50/10 would be 2.6% bigger than it was observed. Even if we are unsure about the impact of total association on high-middle inequality given the inconsistency between share ratios and percentile ratios, at least we can conclude with some confidence that the impact of changes in the total association between spouses' earnings on overall inequality and middle-low inequality is disequalising.

Table 6 Total Association

Measure	Method	1988 observed	2002 observed	2002 counter-factual	Observed Change (Chg1)	Counter-factual Change (Chg 2)	*Contribution(%)*
CV	M1	0.454	0.768	0.747	0.314	0.293	*6.584*
	M4	0.382	0.704	-	0.323	0.277	*14.185*
Gini	M4	0.203	0.379	-	0.176	0.155	*11.919*
high-middle	M4	0.599	0.805	-	0.206	0.161	*21.870*
90/50	M4	1.667	2.242	-	0.575	0.590	*-2.609*
middle-low	M4	4.804	11.491	-	6.687	4.424	*33.841*
50/10	M4	1.543	4.049	-	2.505	1.076	*57.038*

Note: 1.Counterfactual Change (Chg2)=2002 counterfactual-1988 observed
2.Contribution=(1-Chg2/Chg1)*100

How much can the contribution of changes in total association be attributed to its main aspect and how much to the residual aspect? M2 and M3 estimate the impact of changes in the 'main association' (i.e. changes in the association structure between spouses' earnings categories net of structural forces and the composition effects) only, while M4 estimates both. Table 7 shows that the role of changes in the main association is either negligible or even equalising, depending on the methods used. As for overall inequality, both M3 and M4 suggest that the impact of changes in the main aspect of association is tiny, while M2 suggests that it is equalising: without changes in the main aspect of association between spouses' earnings, the increase in the CV of earnings distribution from 1988 to 2002 would be 5% more than it is observed or 9% more if measured by Gini coefficient. I trust more in the estimates of M2 for the reason that the classification scheme of M2 (1000 categories) is far more detailed than that of M3 (6 categories) and M4 (11 categories). In this sense, the impact of changes in the main association is estimated to be nearly zero by M3 and M4 but clearly equalizing by M2 possibly because M3 and M4 fail to preserve sufficient information on changes in the association structure in the data.

Table 7 Main vs Residual Association

Measure	Method	1988 observed	2002 observed	2002 counter-factual	Observed Change (Chg1)	Counter-factual Change (Chg 2)	*Contribution (Main;%)*	*Contribution (Residual;%)*
CV	M2	0.449	0.764	0.779	0.315	0.330	*-4.763*	-
	M3	0.454	0.768	0.767	0.314	0.313	*0.294*	-
	M4	0.382	0.704	-	0.323	0.323	*-0.055*	*14.240*
Gini	M2	0.208	0.384	0.400	0.176	0.192	*-9.098*	-
	M4	0.203	0.379	-	0.176	0.176	*0.068*	*11.851*
high-middle	M2	0.564	0.796	0.806	0.232	0.242	*-4.545*	-
	M4	0.599	0.805	-	0.206	0.195	*5.000*	*16.870*
90/50	M2	1.494	2.171	2.168	0.677	0.673	*0.568*	-
	M4	1.667	2.242	-	0.575	0.575	*0.000*	*-2.609*
middle-low	M2	5.164	13.577	17.640	8.414	12.476	*-48.285*	-
	M4	4.804	11.491	-	6.687	8.253	*-23.406*	*57.247*
50/10	M2	1.578	3.665	5.882	2.087	4.304	*-106.187*	-
	M4	1.543	4.049	-	2.505	2.505	*0.000*	*57.038*

Note: 1.Counterfactual Change (Chg2)=2002 counterfactual-1988 observed
2.Contribution(Relative;%)=(1-Chg2/Chg1)*100
3.Contribution(Absolute;%)=Contribution of Total Association-Contribution of Main Association (E.g. 14.240=14.185-(-0.005);same for the rest)

As measures of overall inequality, CV and Gini may mask trends towards greater or lesser inequality at different parts of the earnings distribution. The results from share ratios and percentile ratios reveal that changes in main association seem to have equalised middle-low inequality. Their impact on high-middle inequality is negligible or slightly disequalising, depending on measure and method used.

Despite the zero or equalising effect of changes in main association, the impact of changes in total association is still disequalising according to results from M1 and M4 in Table 7. This implies that the major source of the disequalising effect of changes in total association comes from the trends in residual correlation between spouses' earnings. Only M4 is able to decompose the effect of changes in total association into main and residual aspects. It shows that almost all of the disequalising impact of the total association comes from trends in residual correlation when trends in overall inequality are considered. As for high-middle inequality, share ratios and percentile ratios lead to estimates of different directions (16.9% vs -2.6%). For middle-low

inequality, however, both show that trends in residual correlation are disequalising, accounting for 57% of the observed increase in the value of middle-low share ratio as well as p50/10.

This decomposition result might be suspected to have underestimated the role of changes in main association for the following reasons. First, M4 classifies earnings into one zero category plus ten deciles, which may not suffice to capture enough information on trends in the association pattern in the original data, or may have twisted the original individual data beyond an acceptable degree that our calculations which are based on the grouped data are likely to be imprecise. Second, in the preferred model of M4, there is only one term for changes in the main aspect of association---changes in the relationship between husbands' earnings categories and the likelihood that wives work. Therefore it is possible that this term is not enough to represent the real changes in the main association, which contains three components in total. The terms for the other two components of changes in main association---changes in the association between dual-earners and the relationship between wives' earnings and the likelihood that husbands work---are excluded because they do not improve the goodness-of-fit of the model, as mentioned earlier. However, it is also possible that I have not found the terms that are specified properly to represent them.

Neither of these two arguments holds true if we recall the results from M2. M2 preserves the rank correlation in 1988 where there are 1000 ranks, which should be detailed enough to capture the main aspect of the original association pattern. Nevertheless, both the CV and Gini of the inequality under the counterfactual that couples in 2002 were matched as their counterparts did in 1988 are slightly higher than their observed level, indicating that the impact of changes in main association is actually equalising or, at least, not disequalising. Therefore, the impact of changes in the residual aspect of association must be disequalising so as to produce the disequalising impact of total association.

What remains is to decompose the contribution of changes in the main association. Table 9 shows the results, where C1, C2 and C3 refer to each of the three components of changes in relative association as explained in Note below the table. M4 finds there is no significant change in C1 and C3, so the log-linear model has no term for them and their contributions are not estimated. For the same reason, the estimates of M4 about the impact of C2 on high-middle and middle-low inequality are equal to those for the impact of the whole main association by M4, which have already been shown in Table 7. Therefore, Table 9 provides estimates on overall inequality (CV) only. It shows that the estimates of M3 differ from those of M4 in some aspects. What they share in common is that the impact of changes in C1 (association between dual-earners) seems negligible. As shown earlier, this is probably due to the fact that there is little change in C1 during this period.

Table 9 Decomposition of The Impact of Main Association

Measure	Method	1988 observed	2002 observed	2002 counter-factual	Observed Change (Chg1)	Counter-factual Change (Chg 2)	*Contribution(%)*
C1	M3	0.454	0.768	0.767	0.314	0.313	*0.203*
C2	M3	0.454	0.768	0.742	0.314	0.288	*8.194*
	M4	0.382	0.704	-	0.323	0.323	*-0.055*
	M4'	0.382	0.705	-	0.323	0.323	*0.081*
C3	M3	0.454	0.768	0.788	0.314	0.334	*-6.421*

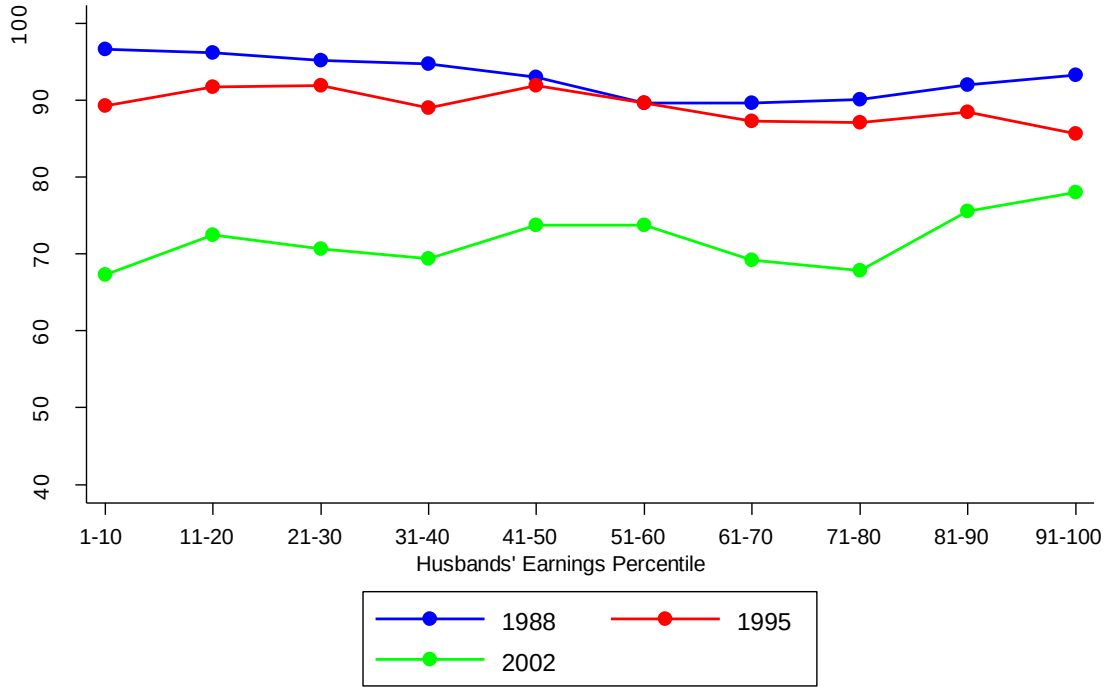
Note: 1. C1=association among dual-earners
2. C2=relationship between husbands' earnings categories and the likelihood that wives work
3. C3=relationship between wives' earnings categories and the likelihood that husbands work
4. M4' is method four where all the cells for C2 are exactly fitted in log-linear models

As for C2, M3 finds that its changes have disequalised the overall inequality: without its changes the increase in overall inequality, if measured by CV, would be 8.2% less than it is observed. However, M4 estimates this impact to be nearly zero (-0.05%). Turn to the statistics by M4 in Table 7, we see that the impact of changes in C2 on high-middle inequality is slightly

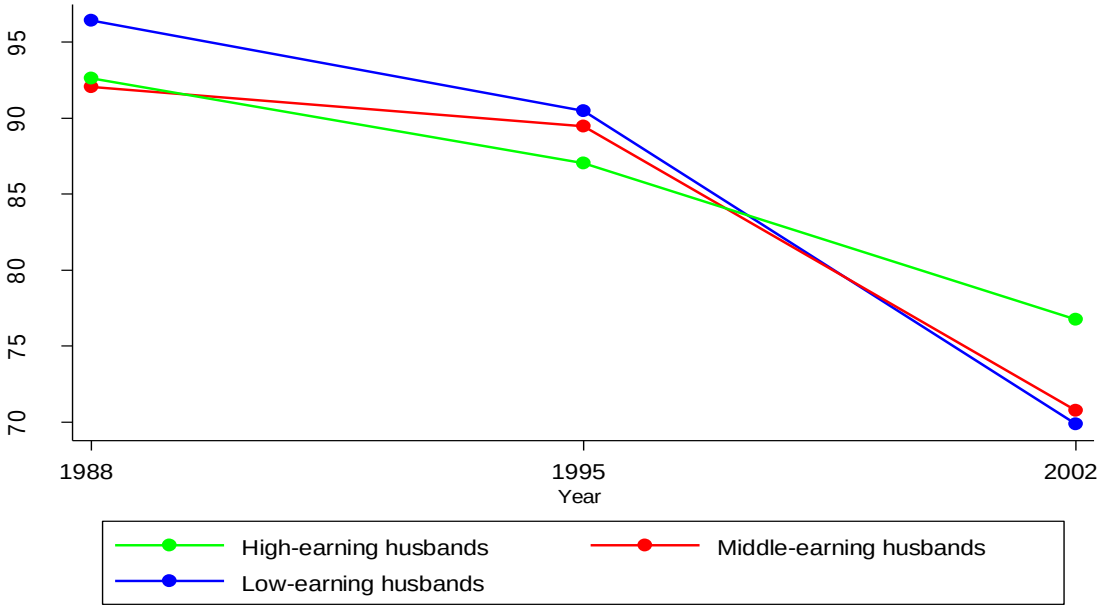
disequalising (5%) when measured by share ratio or zero when measured by p90/50, but its impact on middle-low inequality is equalising (-23%) by share ratio or zero by p50/10. If we trust more in the results of share ratios, it seems to indicate a tendency towards a U-shape relationship between husbands' earnings categories and the likelihood that wives work from 1988 to 2002. However, M3 seems to suggest a tendency towards a more linear relationship that across this period wives of high-earnings husbands are more likely to work or wives or low-earnings husbands are less likely to work. Is this inconsistency due to the fact that the preferred model of M4 uses a quadratic specification of C2 which fails to preserve sufficient information in the original data? To test this possibility, in M4' I use another model where all the cells for C2 are exactly fitted (model 5 in Table 3). It turns out that this brings little change, as can be seen in the forth row of Table 9.

To further understand this point, Figure 1 and 2 are plotted. Figure 1 shows that in 1988 and 1995 there seems to be no apparent association between husbands' earnings deciles and the likelihood that wives work, although wives with richer husbands seemed to be less likely to work in 1988. In 2002, however, some significant changes seem to have taken place that wives whose husbands' earnings are in the top 20% are obviously more likely to work than others. But apart from the contrast between wives with top-20% husbands and wives with other husbands, there seems to be no apparent tendency within wives with middle- or low-earning husbands. Figure 2 tells basically the same story but highlights the tendency that wives with high-earnings husbands are increasingly more likely to work. These two figures are supportive of the estimates of M3 in terms of the impact of changes in the relationship between husbands' earnings categories and the likelihood that wives work. I do not understand why M4 estimates that changes in C2 equalise the share ratio of middle-low inequality (-23%).

% Working Wives by Husbands' Earnings Decile



·centage of wives with non-zero annual earnings by year and husbands' earnings



The impact of changes in C3 differs between M3 and M4, too. M3 finds it equalising (without its change the change of the CV of overall inequality would be 6.4% more than observed) while M4 finds it to be zero. The estimate of M4 seems more plausible given the fact that the labour force participation of husbands is usually not affected by the earnings of wives. M3 finds it to be slightly equalising perhaps because the share of husbands' earnings in family earnings is higher, so any little change in it may produce disproportionately significant outcomes.

In summary, I find that changes in the association between spouses' earnings in urban China are correlated with the increasing inequality of earnings among married couples in urban China from 1988 to 2002. Without its changes, the increase in overall inequality among couples would be 6-14% less than observed, depending on the methods used. The association between spouses' earnings contains two aspects, and almost all of the disequalising effect of changes in the association comes from changes in its main aspect. The disequalising effect seems to be more significant for middle-low inequality than high-middle inequality. The impact of changes in main association is either negligible or, if any, equalizing, again depending on the method and measure used. Among its three components, changes in the association among dual-earners are estimated to have negligible impact on trends in inequality, probably because the association among dual-earners itself changed little across this period. The contribution of changes in the relationship between husbands' earnings categories and the likelihood that wives work are estimated to be disequalising by Method 3 (accounting for 8% of the observed increase in overall inequality) but negligible by Method 4. Method 4 estimates its impact on overall inequality to be almost zero because its changes seem to equalize the high-middle inequality yet disequalise the middle-low inequality, the two of which might have offset each other. The impact of changes in the relationship between wives' earnings categories and the likelihood that husbands work are estimated to be equalizing by Method 3 but again negligible by Method 4. The reason is not clear.

Discussion

This study shows that, in spite of the increasing resemblance in a wide range of socioeconomic traits between spouses during recent decades, husbands and wives in urban China has not become more or less alike in terms of annual earnings, especially among dual-earner couples——the majority of married couples in the context of urban China. As a result, unlike its significant role in the US, changes in the association between spouses' earnings have not significantly contributed to the dramatic increase in family earnings inequality in urban China. Following previous literature that examines family dynamics, this study reveals the insignificance of changes in another family dynamic in accounting for trends in income inequality in China. (Wu 2010; Xie & Zhou 2014). Even if the impact of trends in residual association is taken into account, the total changes in the association between spouses' earnings explain only 6-14% of the observed increase in earnings inequality among couples in urban China. Given the fact that the association between spouses' earnings results from two processes---assortative mating and intrafamily labour division---the small contribution of changes in spouses' earnings association to earnings inequality may be interpreted either as a proof that assortative mating is not important to understanding trends in earnings inequality among couples in urban China (in 1988-2002), or a proof calling for more inquiry that assortative mating might be important in some way, but its disequalising effect may have been largely offset by changes in the pattern of labour division within families. Nevertheless, assortative mating may still be an important mechanism of generating social inequality in other spheres even if it were not essential to earnings inequality.

There are several directions to improve this draft:

1. As noted above, the impact of changes in the association between spouses' earnings measured in this research is not a causal effect but a bivariate correlation. It is likely the estimated size or even direction of the correlation depends on some other variables. For instance, if I use the

non-parametric method (M2) to estimate the level of inequality in 1988 under the counterfactual that only the rank correlation between spouses' earnings had changed to its 2002 level, it turns out that changes in the rank correlation had little impact on trends in inequality, no matter which measure of inequality I use, which differs from the equalising effect found above. One potential explanation is that the baseline distribution upon which I build my counterfactual analysis is different. When changes in earnings rank correlation are the only element that has not changed, my reference distribution upon which the counterfactual analysis is built is the observed earnings distribution in 2002. When changes in correlation are the only element that has changed, however, the baseline is the 1988 distribution. Figure 4 in Appendix 1 shows that the share of dual-earner couples in 1988 is far larger than that in 2002. As a result, when 1988 distribution is used as the baseline and the rank correlation is set to its 2002 level, changes in the association among dual-earner couples would receive more weight than it would do when 2002 distribution is the baseline upon which the rank correlation is set to its 1988 level. As mentioned above, there was actually little change in the association between spouses' earnings ranks among dual-earners across this period. Therefore, the impact of changes in the rank correlation between spouses' earnings is likely to be smaller when changes in the rank correlation reflect more of the changes among dual-earner couples, i.e. when 1988 distribution is the baseline for counterfactual analysis. So one direction of improvement would be to repeat all the procedures above to see how much inequality would change if only the association between spouses' earning had changed.

2. To incorporate single-person household into the sample. It has been found that low-income people (especially men) are more likely to be disadvantaged not only because their partners' incomes are more likely to be low, but also because they are less likely to be in a union (remain unmarried or divorce) in the US (Burtless 1999). If family size is taken into consideration, the inclusion of single-person household will influence the level of inequality because two poor single households have more disequalising effect on inequality than one household of two poor

people. It also makes sense to incorporate single-person household because assortative mating should also includes assortative entry into marital union.

3. To examine the impact of changes in the association between spouses' income on trends in family income inequality, and use CGSS2006 and CGSS2010 to explore the most recent trends because these two datasets contain information on incomes of household head and their spouse. However, this will simplify the analysis and the effect of changes in female labour force participation cannot be identified. Family size adjustment may also be taken into consideration, but I am still not very clear how to deal with this issue using the four methods.

4. I am still not very sure about appropriateness of the definition of earnings I copied from Ding et al. (2009). It takes income tax into account. It differs from the definition used in other studies (Xie & Hannum 1996; Hauser & Xie 2005). The author of Ding et al. (2009) told me in an email that they excluded some cases with extreme values (and she believed that the deletion of these cases affects the value of CV), and they identified the employment status of a person on more than one question while in this draft I rely on a single question in the survey questionnaire. I have not modified my stata codes according to her suggestions yet.

5. I can further repeat all the procedures to compare between 1988 vs 1995 and 1995 vs 2002.

Key References

Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the US income distribution. *European Economic Review*,43(4), 853-865.

Breen, R., & Salazar, L. (2011). Educational Assortative Mating and Earnings Inequality in the United States1. American Journal of Sociology, 117(3), 808-843.

Cancian, M., & Reed, D. (1999). The impact of wives' earnings on income inequality: Issues and estimates. *Demography*, 36(2), 173-184.

Ding, S., Dong, X. Y., & Li, S. (2009). Women's employment and family income inequality during China's economic transition. Feminist Economics, 15(3), 163-190.

Reed, D., & Cancian, M. (2012). Rising family income inequality: The importance of sorting. *Journal of Income Distribution*, 21(2), 3-14.

Schwartz, C. R. (2010). Earnings inequality and the changing association between spouses' earnings. AJS; *American journal of sociology*, 115(5), 1524.

Schwartz, C. R. (2013). Trends and Variation in Assortative Mating: Causes and Consequences. *Annual Review of Sociology*, (0).

Appendix 1 Changes in the association between spouses' earnings

Figure 1

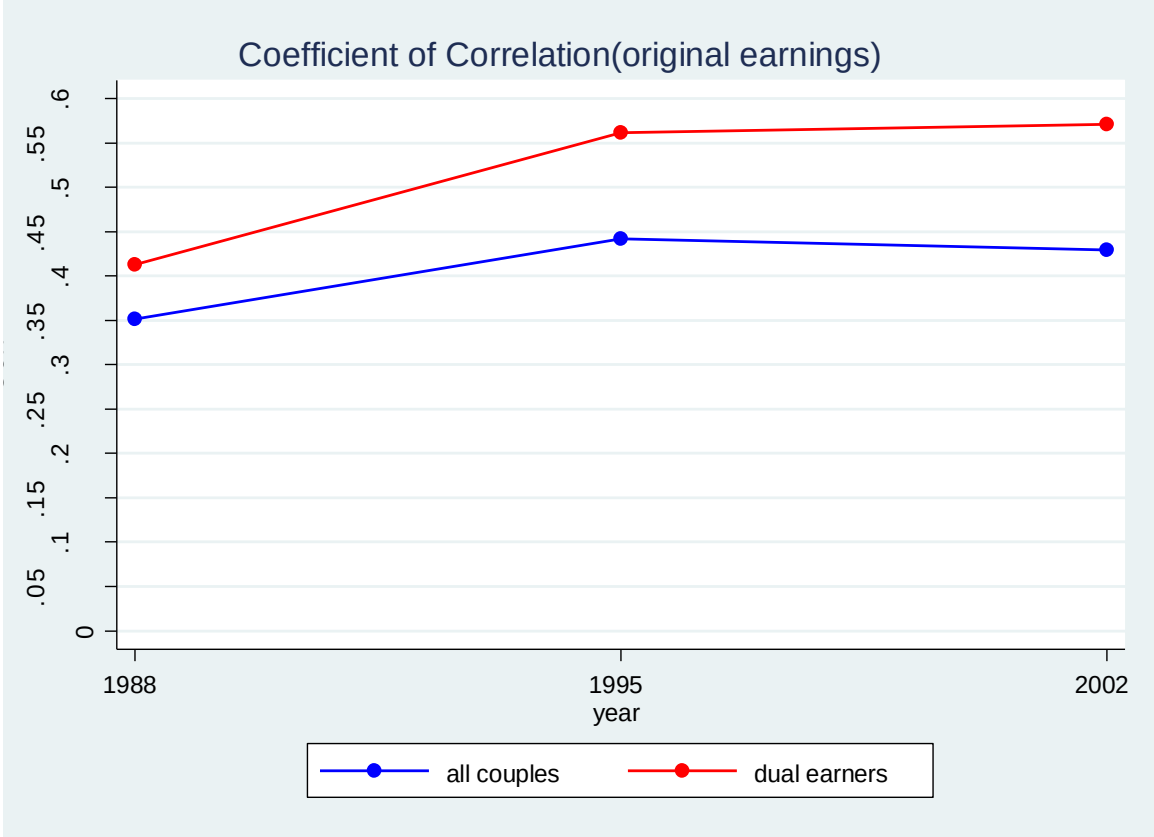
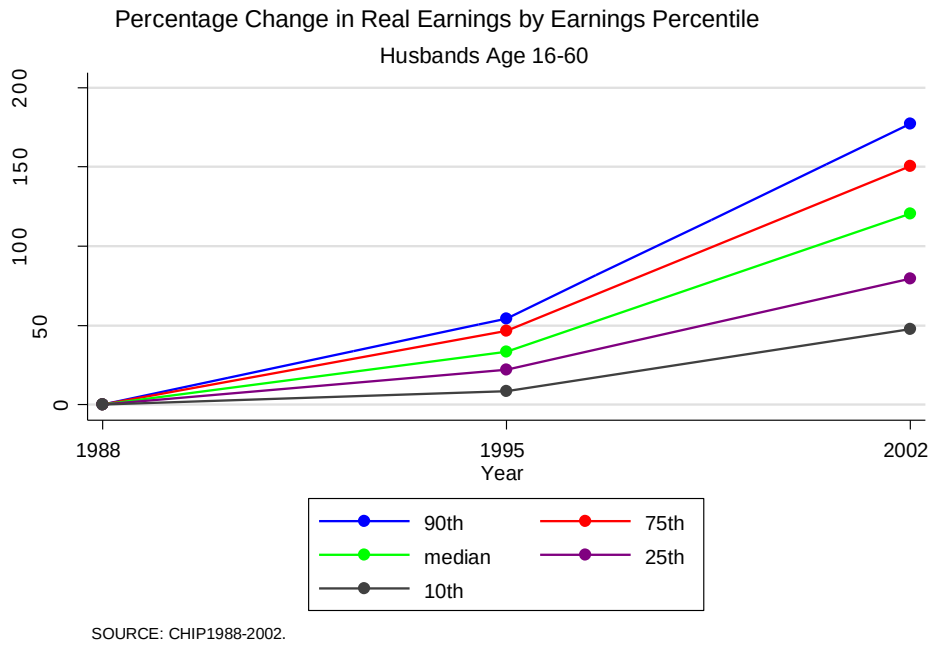


Figure 2



SOURCE: CHIP1988-2002.

Figure 3

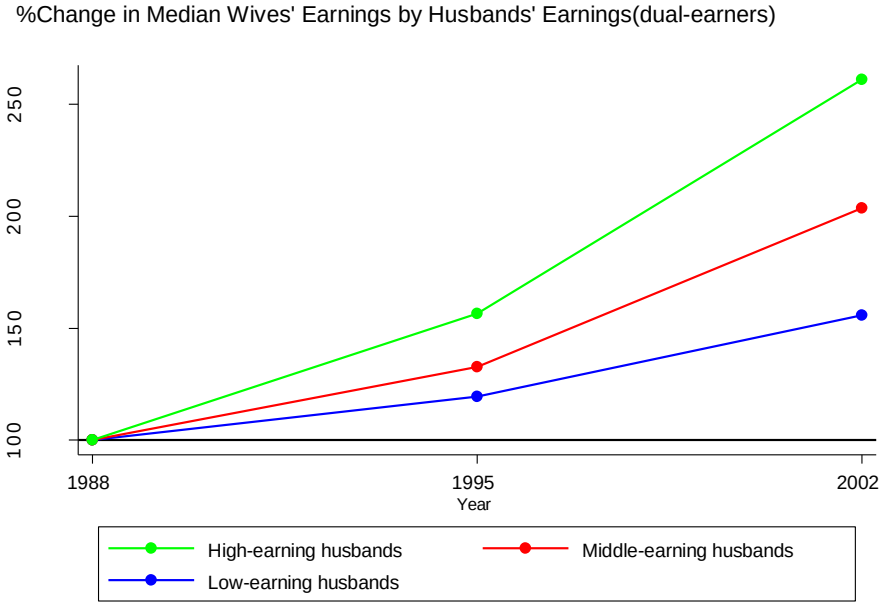
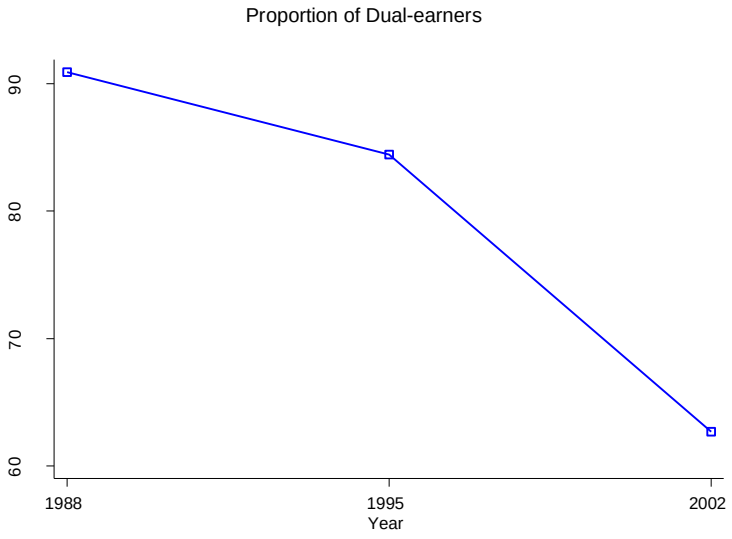


Figure 2 shows that despite the overall growth in husbands' annual earnings during the fourteen years under analysis, husbands in the upper part of the distribution experienced faster increase than those in the middle and lower part. Similarly, the earnings of wives with high-earning husbands (defined as men with top 20% earnings) grew faster than those of wives with middle-earning husbands (defined as men with middle 60% earnings), which also grew faster than those of wives with low-earning husbands (bottom 20%), as illustrated in Figure 3. The combination of these two trends indicates that changes in earnings correlation among dual-earners were shaped by an increasingly enlarged gap between high-earning couples and middle-earning couples as well as between middle-earning and low-earning couples. This suggests a right-skewed change in the earnings distribution of both partners, which increases the value of coefficient of correlation as mentioned in the article.

Figure 4



Appendix 2 Trends in earnings inequality among married couples

Figure 1B

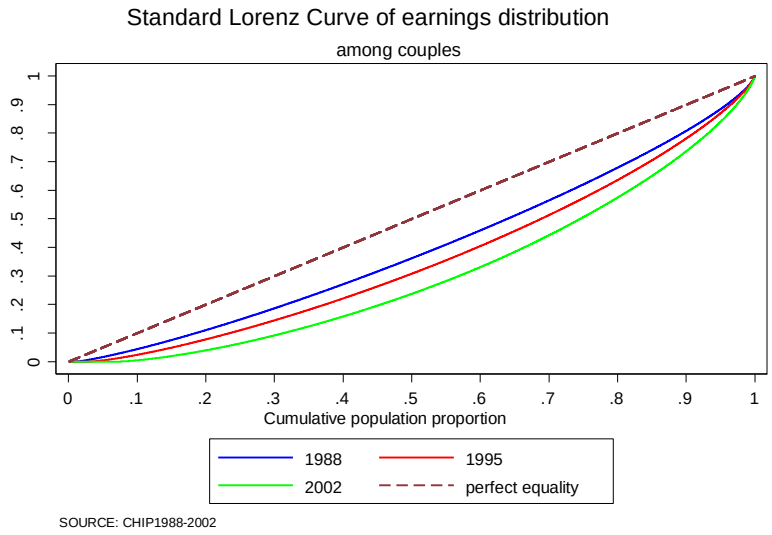
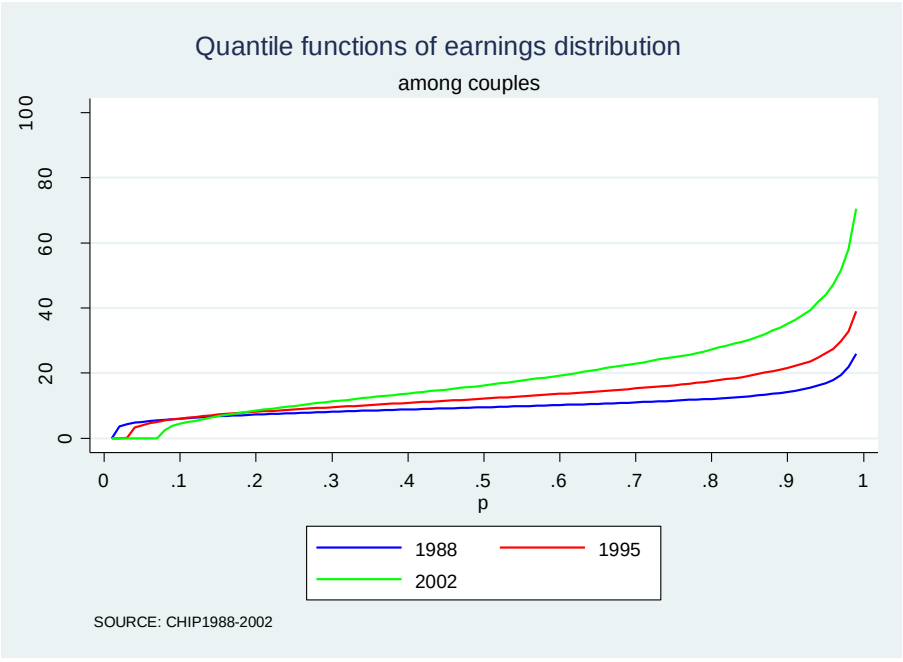


Figure 2B



First, in Figure 1B, changes in the standard Lorenz curves illustrate a clear tendency towards higher overall inequality from 1988 to 2002, which turns out to be driven by substantial deviance at almost all points of the earnings distribution as is shown by the quintile functions in Figure 2B: the real earnings of couples in the bottom 20% in 2002 appear even lower than those in 1988 due to the rise in the share of couples with zero annual earnings, i.e. neither partner work. In fact, while in 2002 the real earnings of couples at the 50^{th} and 90^{th} percentile rose by 71% and 148% compared to 1988, couples' real earnings at the 10^{th} percentile declined by 26%.

I did a set of formal tests on the statistical significance of differences in inequality. Because the data contains observations with zero values, only a few inequality indexes are computed. Nevertheless, all of them confirm our conclusions above that the earnings distribution among couples in urban China becomes more unequal from 1988 to 2002 (see table 1B). Moreover, increases in the middle-low inequality seems more remarkable than increases in high-middle inequality (shown by the change% in table 1B). This is more or less understandable given the dramatic rise in the proportion of couples with zero annual earnings in later years (especially in 2002).

Table1B

Index	1988	1995	diff	P>t	change%
Gini	0.21	0.28	0.08	0.000	36.33
GE(2)	0.10	0.16	0.06	0.000	56.54
p90/50	1.49	1.77	0.28	0.000	18.77
p50/10	1.58	2.02	0.44	0.000	28.18
Index	1995	2002	diff	P>t	change%
Gini	0.28	0.38	0.10	0.000	35.24
GE(2)	0.16	0.29	0.13	0.000	82.89
p90/50	1.77	2.17	0.40	0.000	22.37
p50/10	2.02	3.65	1.63	0.000	80.64

SOURCE:CHIP1988/1995/2002