Report on
Consumption and Social Integration: An Empirical Evidence for Chinese Migrant Workers

The paper investigates the relationship between consumption and social integration (SI, hereafter) of migrant workers in China by a first-hand four-province dataset collected in 2013. Although the dramatic increase of migrant workers in China has been increasingly concerned by the researchers, this paper still fills the gap of the empirical works on the linkage between consumption and SI especially for Chinese migrant workers. The SI is measured by Principal Component Method which covers five major dimensions of SI mentioned by the previous studies. The consumption of Chinese migrant works is measured by three dimensional indicators include consumption level, consumption structure and consumption behavior. The final results show that there is a strong and positive relation between consumption and SI for migrant workers in China. This paper will add contribution to the literature and enrich our knowledge of the link between consumption and SI especially for migrant works in China.

Nevertheless, there is still a lot of work to do before the acceptance for publication in Economics:

(1) In this paper, the authors tried to examine whether the consumption should be regarded as one of the most important determinant variable for SI. However, it's difficult to clarify such determinant effect due to the endogenity issue of the regression models introduced in this paper: is it the increase in Consumption that increases SI or the increase of SI leads to the increase in Consumption?

To solve this endogenity issue, lagged Consumption or indicators reflect the consumption of correspondents in previous periods should be employed in the final regression models.

(2) Has the methodology used to calculate SI mentioned in the section of 3.3 (Page 10) been introduced or employed by any other papers?

In that section, the authors mentioned that $SI_{i,r}$ is "the average score of each component without weight" and "W_r is the weight of component r, which is indicated by the percentile-converted variance contribution rates of each common factor as weights." Why use this two-step measurement to calculate SI? And the details about the process of percentile-converted variance contribution rates are not clear.

(3) Both consumption and income variables are introduced simultaneously in the final regression models from section 4.1 to section 4.3, which might cause the collinearity issue since there should be a high correlation between income and consumption. On the other hand, even if the high correlation could be mitigated to some extent by transferring income variable to dummy variable (Page 13), the

authors should explain the meaning of the regression where both consumption and income are introduced.

(4) The authors test the correlation between Consumption and SI by directly introducing the consumption level in the regressions from section 4.1 to 4.3, and the consumption level changes dramatically across different provinces (say, Henan and Guangdong province). My suggestion is to use scaled consumption level (consumption divided by total income for each correspondent) rather than the exact value of consumption in the final regressions.

(5) In page 12, the authors categorized four types of Consumption and mentioned in the last paragraph; they defined the consumption types as a dummy variable. What I wonder is that how this dummy variable is used in the regression models of section 4.3 in page 18 since the regressions in section 4.3 is based on the sub-sample of different types of Consumption behaviors.

(6) The regional dummy variables should also be introduced into the final regression models. As what authors mentioned that the final data of Chinese migrant workers is collected in four provinces and the differences between these four provinces in Page 7, the regional dummy variables should be employed in the final regressions in order to control the differences across regions.

(7) The number of observations should be the same across different regression models from section 4.1 to 4.3. The authors mentioned that there are 869 effective correspondents in the final sample, but only in section 4.3 that the number of observations is equal to 869 and there are 801 observations in section 4.1 and 4.2. The authors should explain the reason about the loss of number of observations, or trim the dataset and make sure the number of observations to be the same for different models.

(8) The eigenvalues of each factor should be shown in Table 3 (Page 9), and the questionnaire sample should be put in the Appendix.

To conclude, the topic and idea are interesting and meaningful, but the manuscript I received is not ready to publish yet. I'd encourage the authors to improve it.