Appendix II

The 2002 and 2007 CHIP Surveys: Sampling, Weights, and Combining the Urban, Rural, and Migrant Samples

SONG Jin, Terry Sicular, and YUE Ximing*

I. General Remarks

The CHIP datasets consist of urban, rural, and for 2002 and 2007 rural-urban migrant samples. The sizes of these samples are not proportional to their shares in the Chinese national population. Also, their regional distributions differ from those in the population. Consequently, weights are needed in order to make the samples nationally representative.

In this Appendix we discuss the calculation of sample weights that can be used for analysis of the 2002 and 2007 CHIP data. We calculate these weights using data provided by the National Bureau of Statistics (NBS) from the 2000 census and the 2005 1% population sample survey, hereafter called the "2005 mini census." The census and mini census are the most complete available accountings of China's population available. Our sample weights are designed to reflect population shares in the census and the mini census.

We begin with a discussion of the CHIP sampling design and its implications for the calculation of weights (Section II). The calculation of weights requires data on population shares by geographic location and by urban, rural, and migrant classification, which we obtain using data from the 2000 census and the 2005 mini census. Section III discusses the census and mini-census data that we use for this purpose. In order to construct and apply the weights consistently, we must classify the location of residence for all individuals and households and make sure that there is no double counting. The classification of location is discussed in Section IV. The

last section of this Appendix raises some suggestions for implementation of the weights in the analysis of the data.

II. Calculation of Weights

In the CHIP surveys some groups are over-sampled and others are under-sampled relative to their shares in the national population. Here we discuss the construction of weights that can be used to adjust the CHIP samples so that they reflect selection probabilities from the national population.

In past analyses of the CHIP data, a weight adjustment was made only for the rural and urban dimensions. In 2002, for instance, according to the National Bureau of Statistics (NBS) population data China's rural population was 782.4 million and the urban population was 502.1 million, implying that rural and urban shares in the total population were 60.91 percent and 39.09 percent, respectively. The 2002 CHIP urban and rural sample shares, however, were 64.78 percent and 35.22 percent, respectively, so that the rural population was over-sampled and the urban sample was under-sampled. The use of rural-urban weights with the 2002 CHIP data was intended to adjust the shares of the urban and rural samples so that they were identical to the shares of the urban and rural populations in China's national population.

[insert Table AII.1 about here]

In light of questions raised by the project participants and following extensive discussions, we concluded that the sample weights should reflect not only the rural and urban population shares, but also the population shares of the major regions of

China. This conclusion was based on the principle that weights should be determined in light of the approach used to construct the CHIP samples. The CHIP urban and rural sampling methods were designed to represent the conditions in four regions of China—coastal, central, western, and a separate category for large municipalities with provincial status.[1] Table AII.1 provides a list of the provinces and their regional classifications for all rounds of the CHIP survey from 1988 through 2007. In each round, sample provinces were selected from each region so as to reflect the economic characteristics of that region. This was done separately for the urban and rural samples, yielding a total of eight strata.

The CHIP migrant survey was designed to cover the same four regions as the CHIP urban and rural surveys.[2] Including the migrant survey, the 2002 and 2007 CHIP survey datasets comprise twelve strata: rural coastal, rural central, rural western, rural provincial-level municipality; urban coastal, urban central, urban western, urban provincial-level municipality; and migrant coastal, migrant central, migrant western, and migrant provincial-level municipality. Excluding the migrant samples, the CHIP survey datasets comprise eight strata.

We recommend the use of sample weights based on the population shares of these strata. For the 2002 and 2007 rounds, which are the main focus of this Appendix, weights can be applied for analysis of the CHIP rural sample, urban sample, and migrant subsample (keeping only long-term, stable migrants from the migrant sample so as to avoid double-counting—as discussed below), whether they are used separately or in combination. For example, analysis of China's formal urban

population only (i.e., the urban population with a local urban household registration [*hukou*]) would apply weights from the four urban strata to the CHIP urban survey data. Analysis of China's total urban population (including rural-urban migrants) would apply weights from the four urban strata to the CHIP urban survey data and weights from the four migrant strata to the CHIP migrant data (long-term, stable migrants only). Analysis of China's national population would use weights for all twelve strata applied to the respective CHIP rural, urban, and long-term, stable migrant data.

Researchers may wish to use weights that reflect not only regional populations, but also provincial populations. The provinces covered in the CHIP surveys have different population sizes, but the CHIP provincial samples are quite similar in size. Consequently, the probability of being selected from a large sample province is higher than the probability of being selected from a small sample province.

In principle, whether or not the sample weights should reflect provincial populations depends on the way that the samples are constructed within the regions. If regional samples are selected deliberately to ensure that they are representative of the region, then the sample weights need not reflect the provincial population shares. Unfortunately, selection of the CHIP provinces was not done in an entirely transparent manner, and thus it is unclear whether the sample weights should reflect provincial populations. Here we discuss both approaches and provide two sets of weights. Researchers can decide which approach they prefer.

A. Construction of Weights to Reflect Regional Populations

Our sample consists of individuals, each of whom belongs to a stratum. Here "stratum" refers to any of the twelve subgroups discussed above, e.g., urban-coastal, migrant-central, and so on.

The weight w_i^k for individual i in stratum k is equal to:

$$w_i^k = \frac{N^k}{n^k} \quad , \tag{1}$$

where N^k is the population of stratum k, and n^k is the sample size from stratum k.

Thus, for example, if the sample from stratum k contains 1 percent of the population of that stratum, then each sample observation represents 100 people, and the weight for each observation is 100.

Weighting in this way guarantees that the combination of weighted samples from different strata reflects the combined size of those strata in the national population. For example, the size of the combined weighted samples for all urban strata will equal the size of the national urban population. Similarly, the size of the combined weighted samples for all strata in a region, e.g., central China, will equal the size of that region's population.

These weights are a function of the sample and population shares. Let $S^k = N^k/N$ be the share of stratum k in the national population N, and let $s^k = n^k/n$ be the share of the sample from stratum k in the overall sample size n. Then the weight w_i^k for individual i in stratum k can be written as

$$w_i^k = \left(\frac{N^k * n}{n^k * N} \right) * \frac{N}{n} = \frac{S^k}{s^k} * \frac{N}{n} \quad . \qquad (2)$$

In other words, the weight is equal to the stratum's population share divided by the stratum's sample share, scaled up by the ratio of the national population to the total sample size.

Formula (2) is appealing intuitively, as it tells us that the weights depend on whether or not a stratum's share of the population is bigger or smaller than its share of the sample. So, for example, if the share of rural-central China in China's national population exceeds its share in the CHIP sample, then observations from rural-central China would receive a weight greater than one.

Note that *N/n* is the same for all strata. Since regression methods and inequality measures are typically scale-invariant, for most analyses this scaling factor can be dropped and the weights can be calculated simply as the ratio of the population shares to the sample shares.

B. Construction of Weights to Reflect Regional and Provincial Populations

Whether or not weights should also reflect provincial populations depends on how the sample provinces are selected. If the sample provinces and provincial samples within each region are deliberately selected so that their pooled samples are representative of the region, then weights need not reflect the provincial populations.

Such would be the case, for example, if *J* sample provinces were selected out of the *M* provinces in stratum *k*, and these provinces were chosen because their combined populations are representative of the stratum. In this case, drawing a

random sample of n^k individuals from the pooled populations of the sample provinces would be identical to drawing n^k individuals from the entire stratum. The probability of an individual being chosen would be $p_i^k = n^k/N^k$. The same result would apply if random samples were drawn separately for each province, with the sample size for each province n_j^k being proportional to its population size N_j^k. Then the probability of an individual being chosen would be $p_i^k = (n_j^k/N_j^k) * (N_j^k/N^k) = n^k/N^k$. In either case, the sample weights would be identical to those given in (1) above. Therefore, the weights would only need to reflect the regional populations, not the provincial populations.

Suppose instead that the provinces are chosen to be jointly representative of the region, but the size of each provincial sample is not proportional to its population. This is possibly the case for the CHIP samples. Then the weights should reflect that the probability of being selected differs among provinces.

Let N_j^k be the population and n_j^k be the sample size of province j in stratum k. Then the probability of an individual being drawn within a province is $p_i^{j,k} = n_j^k/N_j^k$. The size of the regional sample is the sum of the samples from all the provinces within the region $n^k = \sum n_j^k$. The weight $v_i^{j,k}$ for an individual i located in province j of stratum k can then be written as

$$v_i^{j,k} = \left(\frac{N_j^k}{n_j^k}\right) * \left(\frac{N^k}{n^k}\right) = \left(\frac{N_j^k}{n_j^k}\right) * w^k \quad . \tag{3}$$

One can see that the second term is simply the stratum weight from formula (1), so $v_i^{j,k}$ is equal to the stratum weight w^k times the ratio of the sample province's population to its sample size.

As is the case for the stratum weights w^k shown in (1), the sum of the combined weighted samples for multiple substrata will equal the combined population of those substrata. For example, the size of the combined weighted provincial rural samples will equal the national rural population.

One can restate formula (3) in terms of population and sample shares as follows:

$$v_i^{j,k} = \left(\frac{N_j^k}{n_j^k} \right) * \frac{S^k}{s^k} * \frac{N}{n} \quad . \tag{4}$$

Alternatively, one can see from the first line of (3) that the weights can be written as the ratio of the province's share of the stratum population (S_j^k) to the province's share of the stratum sample (s_j^k):

$$v_i^{j,k} = \frac{S_j^k}{s_j^k} \quad . \tag{5}$$

III. Population Shares: From the 2000 Census and the 2005 Mini Census

Calculation of weights as outlined above requires information about the populations N^k or N_j^k of the different location strata. For 2002 we obtain this information from the Chinese 2000 census and for 2007 we obtain it from the 2005 mini census, which is a 1 percent sample of the national population. Note that the mini census was not constructed entirely according to population shares across provinces. The NBS provides weights that can be used to adjust the mini-census data so that they reflect

more accurately the provincial populations.[3] We use these weights when we calculate population shares from the 2005 mini census.

The census and mini census counted individuals at a specific point in time (for the census, at midnight, October 31, 2000; for the mini census, the night of October 31, 2005). For each individual, the census and mini census contain a location flag as well as other information, such as gender, age, relationship to the household of residence, type of *hukou*, length of time away from the location of the *hukou*, and so forth

We do not have access to the full datasets for the 2000 census and the 2005 mini census; however, the NBS has provided us with randomly selected subsamples. For the 2000 census we have a 0.095 percent sample, and for the 2005 mini census we have a 20 percent subsample. The NBS selected these subsamples using systematic interval sampling, so they should be representative of the full census and the full mini census.

We checked the composition of our subsamples of the census and mini census against the aggregated data from the full census and the full mini census published by the NBS. The subsamples' population shares among provinces, by gender, and by city/town/village are similar to those for the full census and the full mini census.

For calculation of weights we make use of each individual's location (city, town, or village) flag. The location flags in the 2000 census followed certain criteria that were designed to ensure that the census counted stable residents and that people who had moved were not double-counted. An individual was flagged in his or her

location at the time of the census if: (a) he or she was living in and had a *hukou* in that location (including members of households in the location who were not present at the time of the census but had been away for less than six months); or (b) he or she had a *hukou* elsewhere but was living in that location at the time of the census and had been living there for more than six months.[4]

The 2005 mini census used a different approach. All individuals were flagged in their location at the time of the mini census. In addition, individuals who were members of households in a location and had a *hukou* in that location but were away (*waichu renkou*) at the time of the mini census were flagged as residents of that location. This approach might lead to some double-counting of individuals who were away from their households at the time of the mini census.[5]

IV. Classification of Location

In order to construct weights we need to classify individuals according to their location of residence into the different strata. This classification must be done consistently for all datasets used to construct weights, that is, for the 2000 census, the 2005 mini census, and the 2002 and 2007 CHIP urban, rural, and migrant samples. As each location is either urban (including cities and towns) or rural (villages), the consistent classification of individuals by location ensures consistent classification of individuals as urban or rural. The classification is applied to all individuals, including migrants. Migrants who, according to the classification criteria, are classified as residents of a city or town will be counted as urban; those classified as

residents of a village will be counted as rural.

The criteria we adopt for classification of location are those used by the NBS in its annual rural and urban household surveys. The CHIP rural and urban household survey samples are subsets of the NBS rural and urban household surveys, therefore using the same criteria is practical. The NBS criteria consider not just the location and length of residence, but also the strength of economic ties between the individuals and the households.

The NBS (and CHIP) urban and rural survey samples consist of households and their members. An individual is counted as a resident in a location if he or she is a member of a household in that location and if he or she is usually living in the household or has lived in that household for six months or more during the survey year. An individual who is not usually living in the household or who is away from the household for more than six months is counted as a resident if most of his or her income is returned to the household, or if he or she maintains a close economic relationship with the household. Individuals who do not satisfy these criteria are not counted as residents of the location.

A. Reclassifications of the Census and Mini-Census Samples

The criteria used to flag location in the 2000 census and 2005 mini census are different from those used in the NBS household surveys, so we must reclassify the individuals in the census and mini census before constructing the population shares and sample weights.

The most important difference is in the treatment of individuals who are away from their households for more than six months but maintain an economic relationship with the household. The census and mini census count these individuals in their place of residence; we must reclassify them in the location of their households of origin.

The census and mini census do not contain information about the strength of an individual's economic relationship with his or her household of origin, but they contain information about marital status and about whether the individuals are living with their spouses. We use this information as a proxy for the strength of their relationship with the household of origin. If an individual with a non-local *hukou* is married and not living with his or her spouse, we consider that person as having a significant economic relationship with his or her household in the location of the *hukou*. We consider such individuals to be unstable migrants. If an individual with a non-local *hukou* is not married (single, divorced, or widowed), or is married and is living together with his or her spouse, then we consider that person as not having a strong economic relationship with his or her household in the location of the *hukou*. We consider such individuals to be stable migrants.

We must also carry out some additional reclassifications of individuals in the 2005 mini census because the approach used to flag location in the 2005 mini census is different than that used in the 2000 census. For consistency with the census and the NBS household surveys, we reclassify individuals who have lived in the location at the time of the mini census for less than six months in the place of their *hukou*.[6]

In order to carry out these location reclassifications, we examine all individuals in our subsamples of the census and mini census. We accept the flagged location and do not reclassify individuals who satisfy the following conditions:

1. They hold a local *hukou* (regardless of whether the local *hukou* is agricultural or non-agricultural) and (a) they are currently living in the location, or (b) they are absent but they are members of local households and have been away for less than six months,
2. They do not hold a local *hukou* but have been living in the flagged location for more than six months and are either (a) single, divorced, or widowed, or (b) married and living with spouse.

All other individuals are reclassified as a resident in the province of their *hukou*. In other words, all individuals who do not hold a local *hukou* and have been living in the flagged location for less than six months are reclassified, as are all individuals who do not have a local *hukou* and have been living in the flagged location for more than six months, are married, and are not living with their spouses.

Individuals who are reclassified back to the province of their *hukou* will be designated as rural or urban, based on whether they have an agricultural or nonagricultural *hukou*. If they have an agricultural *hukou*, they are reclassified as a rural resident of the province of their *hukou*; if they have a nonagricultural *hukou*, they are reclassified as an urban resident of that province.

This reclassification scheme effectively treats temporary migrants and long-term, unstable migrants as residents of the place of their *hukou*. Migrants who are long

term and stable are not reclassified. Note that reclassification can occur for any type of migrant, including urban-urban, rural-rural, urban-rural, or rural-urban. Rural-urban migrants, however, are of particular interest and are the most numerous.

[insert Table AII.2 around here]

Table AII.2 gives a summary of the 2000 census and 2005 mini-census samples before and after reclassification. For the 2000 census, reclassifications were mainly confined to individuals who lived in the location for more than six months and were married but not living with a spouse. There were more reclassifications for the 2005 mini census because in the mini census individuals who have lived in the location for less than six months were also reclassified.

B. Reclassifications of the CHIP Survey Samples

Because we have adopted the location criteria used in the NBS urban and rural household surveys, and because the CHIP urban and rural samples are drawn from the NBS household surveys, we do not need to reclassify individuals in the CHIP urban and rural survey samples. We treat all individuals in the CHIP rural sample as residents in their given rural locations, and all individuals in the CHIP urban sample as residents in their given urban locations.[7]

The NBS rural surveys treat individuals who live in their rural households of origin most of the time, or who live away from the household for more than six months but maintain a close economic relationship with the household as members of the rural households. Short-term, unstable, rural-urban migrants are therefore

counted in the rural survey. The problem of under-representation of migrants, then, occurs mainly for longer-term rural-urban migrants who do not maintain a close economic relationship with their rural households.

This group of migrants is included in the CHIP migrant surveys. The CHIP migrant surveys also include other types of individuals. These surveys are samples of individuals with agricultural *hukou* who live in urban areas, including not only long-term, stable rural-urban migrants, but also individuals with local agricultural *hukou*, short-term rural-urban migrants, and long-term rural-urban migrants who maintain a close economic relationship with their rural households of origin. For the purpose of calculating weights, we need to drop these latter types of individuals, as they are already included in the NBS and CHIP rural surveys.

On this basis, we only keep individuals in the CHIP migrant surveys who have non-local *hukou* and satisfy the following criteria:[8]

1. They have been living in the urban location for more than six months and they are single, divorced, or widowed, or
2. They have been living in the urban location for more than six months and they are married and living with their spouses.

We call these individuals long-term stable migrants. Individuals who have been living in the urban location for less than six months, or for more than six months but are married and not living with their spouses, are dropped. We call these individuals short-term or long-term unstable migrants. Individuals who have local agricultural *hukou* are also dropped.

Table AII.3 shows the number (and percentage) of individuals in the 2002 and 2007 CHIP migrant surveys that satisfy the above criteria for long-term, stable migrants. It also shows the number of individuals in the migrant surveys who belong to other categories. Note that the different compositions of the 2002 and 2007 migrant samples reflect in part the differences in the sampling methods used to construct the migrant samples in the two years. The use of neighborhood committees as the sampling frame in 2002 led to a higher proportion of long-term stable migrants and individuals with local agricultural *hukou*.

[insert Table AII.3 around here]

We have created a variable *catg* for the 2002 and 2007 migrant datasets that identifies individuals as long-term, stable migrants according to these criteria. The Stata data files *mcatg02.dta* and *mcatg07.dta* contain this variable and ID variables to facilitate merging with the CHIP migrant survey datasets (available on request from the authors). The variable *catg* can be used to keep or drop observations. When calculating weights and using the migrant data in combination with the CHIP urban and rural datasets, observations with *catg* = 2 satisfy the criteria for long-term, stable migrants and should be kept; all other observations should be dropped.[9]

V. Implementation of Weights

Tables AII.4 and AII.5 contain the numbers of individuals in each of the twelve strata and their component provinces in our subsamples of the 2000 census and the 2005 mini census, after reclassification. Researchers can use these numbers as values for

N^k or N_j^k in the calculation of weights. Sample sizes for each of the strata S^k and its component provinces S_j^k will vary depending on the sample used in the analysis, so researchers will calculate these based on the set of observations used in their analyses.

[insert Tables AII.4 and AII.5 around here]

The numbers in Tables AII.4 and AII.5 are appropriate for calculation of weights in analyses at the individual or per capita level. Analyses at the household level should use weights calculated using counts of households, as the number of individuals per household differs among the strata. Tables AII.6 and AII.7 give the counts of households in our subsamples of the 2002 census and 2007 mini census for each stratum and its component provinces. Researchers can use these numbers as the population frequencies N^k or N_j^k for calculation of the household-level weights. Sample counts of households will depend on the observations actually covered in the analysis and thus should be calculated by the researcher accordingly.

[insert Tables AII.6 and AII.7 around here]

This Appendix discusses the calculation of weights based on the geographic distribution of the population among regions and provinces, as well as among urban, rural, and migrant groups. Some researchers may be interested in different subdivisions of the population, for example, between Han and minority groups, or among education groups or age cohorts. Researchers who are analyzing such subgroups will wish to construct weights to ensure that the results are representative of those subgroups. In these cases, one can combine weights based on the regional strata discussed here with weights based on the populations and sample sizes of the

subgroups of interest. For example, Chapter 5 in this volume by Knight, Sicular, and Yue about intergenerational educational mobility uses weights based on age cohorts. Similarly, an analysis of the differences between Han and minority groups might use weights that reflect the sizes of the Han and minority populations in each stratum.

So as to avoid double-counting, in our classification of individuals (and households) by location we have chosen to drop individuals in the CHIP migrant survey who have local agricultural *hukou* because such individuals are also included in the CHIP urban sample. Some researchers, however, may wish to add these urban residents to the CHIP urban sample rather than dropping them. If so, the weights will need to be adjusted accordingly. Similarly, we have dropped short-term and unstable migrants from the CHIP migrant survey because they are also included in the CHIP rural sample. Some researchers may wish instead to add this group to the CHIP rural sample, in which case once again the weights will need to be adjusted accordingly.

Finally, as discussed in Appendix I, we note that for 2007 not all variables are available for the full rural and urban CHIP samples. The sample size will therefore depend on which variables are being used and the number of individuals or households for which they are available. Researchers will therefore need to pay close attention to their sample sizes and recalculate the weights accordingly.

Table AII.1. *Provinces and their regional classifications in the CHIP samples, 1988 through 2007*

province	province code	region	1988		1995		2002			2007				
			rural	urban	rural	urban	rural	urban	migrant	rural CHIP	urban CHIP	migrant CHIP	rural NBS	urban NBS
Beijing	11	1	*	*	*	*	*	*	*				*	*
Tianjin	12	1	*											
Shanghai	31	1	*								*	*		
Hebei	13	2	*		*		*			*				
Liaoning	21	2	*	*	*	*	*	*	*				*	*
Jiangsu	32	2	*	*	*	*	*	*	*	*	*	*		
Zhejiang	33	2	*		*		*			*	*	*		
Fujian	35	2	*										*	*
Shandong	37	2	*		*		*							
Guangdong	44	2	*	*	*	*	*	*	*	*	*	*		
Hainan	46	2	*											
Shanxi	14	3	*	*	*	*	*	*	*				*	*
Jilin	22	3	*		*		*							
Helongjiang	23	3	*											
Anhui	34	3	*	*	*	*	*	*	*	*	*	*		
Jiangxi	36	3	*		*		*							
Henan	41	3	*	*	*	*	*	*	*	*	*	*		
Hubei	42	3	*	*	*	*	*	*	*	*	*	*		
Hunan	43	3	*		*		*						*	*
Inner	15	4	*											

Mongolia														
Guangxi	45	4	*				*							
Chongqing	50	4			(*)	(*)	*	*	*	*	*	*		
Sichuan	51	4	*		*	*	*	*	*	*	*	*		
Guizhou	52	4	*		*		*							
Yunnan	53	4	*	*	*	*	*	*	*				*	*
Tibet	54	4												
Shaanxi	61	4	*		*		*							
Gansu	62	4	*	*	*	*	*	*	*				*	*
Qinghai	63	4	*											
Ningxia	64	4	*											
Xinjiang	65	4					*							

Notes:

1. * indicates that the province is in the sample. For 2007 the columns denoted by CHIP and NBS indicate whether the provinces are covered by the CHIP questionnaire and/or by the supplementary dataset supplied by NBS.
2. The geographic regions are: (1) large municipalities with provincial status, (2) coastal regions, (3) central regions, and (4) western regions.
3. In the original 1988 CHIP sampling frame, Hebei was classified as part of the central region, but its official NBS classification is coastal. Hence, here we have adopted the NBS classification.
4. Chongqing became a separate province in 1997. It was included in the urban Sichuan sample starting in 1995. For consistency over time, and because Chongqing is less urbanized and does not resemble the other large municipalities, we classify Chongqing in the western region.

Table AII.2. *Summary of the 2000 census and 2005 mini census samples before and after reclassification*

	Original		Reclassified Out		Reclassified In		Missing data (dropped)	After Reclassification			
	number	% of population	to rural	to urban	from rural	from urban		number	% local	% migrant	% of population
2000 Census											
urban	432315	36.6%	8293	0	191	0	2380	421833	93.1%	6.9%	35.8%
rural	747795	63.4%	0	191	0	8293	223	755674	100%	0%	64.2%
2005 Mini Census											
urban	1147410	43.7%	25598	5914	926	5914	4642	1118167	92.9%	7.1%	43.7%
rural	1417005	55.3%	7769	926	7769	25598	1137	1440825	56.3%	100%	56.3%

Notes:

1. The numbers for the 2005 mini census are weighted by *power_2*.
2. In principle, the original number plus the numbers “reclassified in” minus the numbers “reclassified out” and missing should equal the post-reclassification number. This is true for the 2000 census numbers, but small discrepancies exist for the 2005 mini-census numbers due to weighting. Without weighting, the equality holds.

Table AII.3. *Composition of the CHIP migrant samples, 2002 and 2007*

Category	2002	2007
(1) Local agricultural *hukou*	1,938 (36.4%)	1,806 (21.4%)
(2) Long-term stable	2,976 (55.9%)	5,303 (62.8%)
(3) Short-term and long-term unstable	278 (5.2%)	1,289 (15.3%)
(4) Missing	135 (2.5%)	98 (1.2%)
Total	5,327 (100%)	8,446 (100%)

Note: This table gives the number of individuals; percentages of the migrant sample for that year are shown in parentheses.

Table AII.4. *Population frequency by stratum, 2000 (individuals in the 0.095 percent subsample of the 2000 census)*

Province code	Province name	Region code	Urban Locals	Stable Long-term Migrants	Rural Locals
11	Beijing	1	8,597	984	3,232
12	Tianjin	1	6,369	289	2,800
31	Shanghai	1	11,796	1,404	1,871
	Subtotal	1	26,762	2,677	7,903
13	Hebei	2	16,202	670	48,964
21	Liaoning	2	21,479	824	18,727
32	Jiangsu	2	26,866	1,659	41,626
33	Zhejiang	2	18,647	1,966	23,717
35	Fujian	2	11,311	1,230	18,971
37	Shandong	2	31,845	1,128	56,274
44	Guangdong	2	32,291	8,171	36,769
46	Hainan	2	2,617	196	4,287
	Subtotal	2	161,258	15,844	249,335
14	Shanxi	3	10,382	451	20,946
22	Jilin	3	11,935	359	12,530
23	Heilongjiang	3	16,473	760	16,079
34	Anhui	3	14,395	514	41,806
36	Jiangxi	3	9,461	310	25,650
41	Henan	3	19,552	785	69,598
42	Hubei	3	20,500	951	30,636
43	Hunan	3	14,974	650	42,325
	Subtotal	3	117,672	4,780	259,570
15	Inner Mongolia	4	8,769	765	12,903
45	Guangxi	4	10,642	695	30,713
50	Chongqing	4	8,801	313	17,844
51	Sichuan	4	19,112	841	55,774
52	Guizhou	4	7,316	597	26,328
53	Yunnan	4	7,953	981	31,672
54	Tibet	4	325	77	1,967
61	Shaanxi	4	10,022	371	23,532
62	Gansu	4	5,425	286	18,585
63	Qinghai	4	1337	120	3,188
64	Ningxia	4	1,606	122	3,718
65	Xinjiang	4	5,757	607	12,642
	Subtotal	4	87,065	5,775	238,866
Total			392,757	29,076	755,674

Table AII.5. *Population frequency by stratum, 2005 (individuals in the 20 percent subsample of the 2005 mini census)*

Province code	Province name	Region code	Urban Locals	Stable Long-term Migrants	Rural Locals
11	Beijing	1	20,476	3,085	4,754
12	Tianjin	1	13,408	1,355	5,186
31	Shanghai	1	24,010	4,687	3,602
	Subtotal	1	57,894	9,127	13,542
13	Hebei	2	46,357	1,347	90,331
21	Liaoning	2	46,773	2,112	32,848
32	Jiangsu	2	72,030	6,767	65,983
33	Zhejiang	2	43,135	7,548	42,212
35	Fujian	2	28,168	4,982	35,009
37	Shandong	2	80,865	3,029	95949
44	Guangdong	2	86,997	21,147	72,754
46	Hainan	2	8,211	598	8,068
	Subtotal	2	412,535	47,530	443,153
14	Shanxi	3	27,336	1,020	39,042
22	Jilin	3	26,608	845	25,644
23	Heilongjiang	3	40,744	1,674	32,127
34	Anhui	3	51,554	1,379	85,144
36	Jiangxi	3	31,078	590	52,585
41	Henan	3	59,245	1,180	129,229
42	Hubei	3	47,604	2,080	64,398
43	Hunan	3	43,862	1,754	81,831
	Subtotal	3	328,032	10,522	510,001
15	Inner Mongolia	4	24,114	2,312	21,754
45	Guangxi	4	28,168	1,259	58,636
51	Sichuan	4	24,945	662	31,407
50	Chongqing	4	49,475	1,857	116,263
52	Guizhou	4	18,709	1,100	55,322
53	Yunnan	4	25,777	2,136	62,381
54	Tibet	4	1,559	88	4,754
61	Shaanxi	4	30,038	1,092	45,526
62	Gansu	4	15,383	446	38,466
63	Qinghai	4	3,742	215	6,483
64	Ningxia	4	4,573	255	7,203
65	Xinjiang	4	13,512	1,108	25,932
	Subtotal	4	239,996	12,530	474,128
Total			1,038,458	29,076	1,440,825

Table AII.6. *Population frequency by stratum, 2000 (households in the 0.095 percent subsample of the 2000 census)*

Province code	Province name	Region code	Urban Locals	Stable Long-term Migrants	Rural Locals
11	Beijing	1	2,914	330	880
12	Tianjin	1	2,131	82	773
31	Shanghai	1	4,103	514	614
	Subtotal	1	9,148	926	2,267
13	Hebei	2	4,723	206	13,165
21	Liaoning	2	7,041	246	5,499
32	Jiangsu	2	8,443	538	12,111
33	Zhejiang	2	6,104	679	7,495
35	Fujian	2	3,274	381	4,931
37	Shandong	2	9,712	287	16,944
44	Guangdong	2	8,758	2,160	8,460
46	Hainan	2	699	60	982
	Subtotal	2	48,754	4,557	69,587
14	Shanxi	3	3,095	116	5,497
22	Jilin	3	3,789	102	3,540
23	Heilongjiang	3	5,389	228	4,585
34	Anhui	3	4,399	128	11,697
36	Jiangxi	3	2,776	69	6,931
41	Henan	3	5,683	176	18,267
42	Hubei	3	6,150	273	8,632
43	Hunan	3	4,789	189	12,264
	Subtotal	3	36,070	1,281	71,413
15	Inner Mongolia	4	2,877	206	3,723
45	Guangxi	4	3,136	196	7,887
50	Chongqing	4	2,944	89	5,627
51	Sichuan	4	6,278	228	16,504
52	Guizhou	4	2,126	159	6,695
53	Yunnan	4	2,533	313	7,913
54	Tibet	4	115	29	404
61	Shaanxi	4	3,042	116	6,203
62	Gansu	4	1,675	87	4,321
63	Qinghai	4	432	31	706
64	Ningxia	4	505	34	855
65	Xinjiang	4	1,773	191	3,073
	Subtotal	4	27,436	1,679	63,911
Total			121,408	8,443	207,178

Note: Includes collective households for migrants, but not for urban and rural locals.

Table AII.7. *Population frequency by stratum, 2005 (households in the 20 percent subsample of the 2005 mini census)*

Province code	Province name	Region code	Stable Urban Locals	Stable Long-term Migrants	Stable Rural Locals
11	Beijing	1	13,413	2,025	2,939
12	Tianjin	1	24,032	1,798	5,941
31	Shanghai	1	24,199	5,287	3,864
	Subtotal	1	61,644	9,110	12,744
13	Hebei	2	11,190	335	19,923
21	Liaoning	2	16,966	741	10,623
32	Jiangsu	2	17,522	1,853	15,906
33	Zhejiang	2	12,994	2,606	12,923
35	Fujian	2	8,531	1,851	10,228
37	Shandong	2	23,875	833	27,360
44	Guangdong	2	61,217	13,158	44,665
46	Hainan	2	5,830	420	4,892
	Subtotal	2	158,125	21,797	146,520
14	Shanxi	3	18,449	677	23,593
22	Jilin	3	15,042	423	12,505
23	Heilongjiang	3	16,014	550	10,498
34	Anhui	3	12,116	308	18,841
36	Jiangxi	3	9,288	145	15,114
41	Henan	3	11,142	195	22,658
42	Hubei	3	15,712	577	19,977
43	Hunan	3	13,159	512	23,123
	Subtotal	3	110,922	3,387	146,309
15	Inner Mongolia	4	10,833	853	8,947
45	Guangxi	4	8,585	354	16,378
51	Sichuan	4	11,467	253	13,815
50	Chongqing	4	12,524	412	27,718
52	Guizhou	4	6,411	334	1,050
53	Yunnan	4	14,221	1,218	36,763
54	Tibet	4	1,318	103	2,915
61	Shaanxi	4	17,008	589	25,270
62	Gansu	4	10,714	293	20,541
63	Qinghai	4	4,658	280	6,044
64	Ningxia	4	3,956	181	4,491
65	Xinjiang	4	6,132	503	8,963
	Subtotal	4	107,827	5,373	188,895
Total			438,518	3,9667	494,468

Note:

Includes collective households for migrants, but not for urban and rural locals.

*The need for careful attention to weights was raised by Samuel L. Myers, Jr., Ding Sai, and Li Shi in "Sample Weights and the Analysis of Per Capita Income: The Case of CHIPs," presented at the CHIP workshop in May 2009. This note builds upon their work. We thank LI Shi for contributing key ideas and for making available the subsamples of the 2000 census and the 2005 1% population survey for use in calculating the weights. This work was supported in part by the Roy Wilkins Center for Human Relations and Social Justice, Hubert H. Humphrey Institute of Public Affairs, University of Minnesota.

[1] See Li Shi, Luo Chuliang, Wei Zhong, and Yue Ximing, "Appendix: The 1995 and 2002 Household Surveys: Sampling Methods and Data Description," in B.A. Gustafsson, S. Li, and T. Sicular, eds., *Inequality and Public Policy in China*, Cambridge: Cambridge University Press, 2008, for an explanation of the sample selection. The geographic regions used to construct the CHIP sample frame are (1) large municipalities with provincial status (Beijing, Tianjin, and Shanghai are treated together as a separate geographic area; (Chongqing is treated as part of western China for consistency with earlier rounds of the survey, when it was included in Sichuan), (2) coastal China (Hebei, Liaoning, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong, and Hainan); central China (Shanxi, Jilin, Heilongjiang, Anhui, Jiangxi, Henan, Hubei, and Hunan); and western China (Inner Mongolia, Guangxi, Chongqing, Sichuan, Guizhou, Yunnan,

[2] The migrant survey for 2002 was carried out in the same twelve provinces as the urban survey, but it covered fewer cities within each province. The migrant survey for 2007 was carried out in nine provinces that were also covered in the 2007 urban survey, but in total the 2007 urban survey covered sixteen provinces.

[3] The weight variable's name is *power_2*. The data contain a value of this variable for each individual, taking 590 different values ranging from .082149 to 2.454594.

[4] People present in a location at the time of the census were also flagged in that location if (a) they had lived in the location for less than six months but had left the place of their *hukou* for more than six months, or (b) they had no *hukou* but they were living or used to live in that location (e.g., newborns, or people studying abroad temporarily). The details of how the locations were flagged in the 2000 census can be found at http://www.stats.gov.cn/tjsj/ndsj/renkoupucha/2000pucha/html/append5.htm, accessed October 11, 2010.

[5] The details of how the locations were flagged in the 2005 mini census can be found *2005 nian quanguo 1% renkou chouyang diaocha ziliao* (Tabulation on the 2005 National Sample Survey of 1 Percent of the Population), Beijing: Zhongguo tongji chubanshe, 2008, p. 833.

[6]The mini census contains information about place of *hukou* (province), type of *hukou* (agricultural or non-agricultural), and length of time away from the place of the *hukou*. We use this information to carry out this reclassification. Information about the length of time is given by the answer to the question (R8) "How much time since he/she left the place of his/her *hukou* registration" (*likai hukou dengji di shijian*). This is slightly different from asking how long the individual has lived in the current location. For example, it is possible that a migrant may have left the original place of the *hukou* a long time ago, first going elsewhere and only recently moving to the current place of residence. We have no information about where the individuals have resided since leaving the place of their *hukou* registration. We assume that individuals who have been away from the place of their *hukou* for six months or more have been living for the last six months in their current place of residence.

[7] We checked the CHIP urban surveys and in fact found some individuals who have non-local rural *hukou*, but these proportions are very small—less than 1 percent of the total observations.

[8] The 2007 CHIP migrant survey contains the question "How many months have you stayed outside your hometown for work or business?" (*Zuijin 12ge yue nei, zai waichu wugong jingshangde yigong shenghuole jige yue?*). The 2002 CHIP migrant survey contains a similar question, "How many months did you stay in an urban area in 2002" (*Zai 2002 nian nin zonggong zai chengzhen juzhu shijian duoshao [yue]?)* We use the answers to these two questions to determine the individual's migration time.

[9] The variable *catg* takes a value of "1" if the individual has a local agricultural *hukou*, "2" if she is a long-term stable migrant, and "3" if she is a short-term or unstable long-term migrant. A missing value indicates that the individual cannot be identified as a member of any of these three groups. We drop individuals if they have a missing value. Researchers can follow our approach and drop them, or they can use other information in the datasets to classify them and include them in their calculations.