ITC Germany Survey Waves 1 (2007) and 2 (2009) Technical Report

Prepared by C. Boudreau

Dept. of Statistics & Actuarial Science and Data Management Core (DMC) of ITC Project

and

M. Yan

Data Management Core (DMC) of ITC Project and Propel Centre for Population Health Impact

University of Waterloo

Updated July 6, 2010

Construction and Use of Sampling Weights for the International Tobacco Control (ITC) Germany Survey

C. Boudreau[1,2] & M. Yan[2,3]

Updated: July 6, 2010 (Waves 1–2)

This technical report details the sampling design and weight construction for waves 1–2 of the International Tobacco Control (ITC) Germany Survey. Section 1 describes the sampling design of the ITC Germany Survey, and section 2 details the construction of the sampling weights for wave 1 (section 2.2) and wave 2 (section 2.3).

1 Sampling Design

The ITC Germany Survey is a prospective longitudinal study, and its sampling design was chosen to yield representative random samples of adult smokers and non-smokers residing in that country. Respondents were first interviewed in July–November 2007 (wave 1), with a follow-up interview in July-October 2009. Both interviews were conducted by the Institut fuer Demoskopie Allensbach, and used computer assisted telephone interviews (CATI).

To qualify for the study respondents must be 18 years old or more. Those that have smoked more than 100 cigarettes in their life and smoked at least once in the 30 days prior to recruitment were considered to be smokers, whereas the others were considered to be non-smokers.

[1]Dept. of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
[2]Data Management Core (DMC) — ITC Project, University of Waterloo
[3]Propel Centre for Population Health Impact, University of Waterloo

1.1 Wave 1

Like most ITC surveys in western countries, the ITC Germany Survey follows a stratified random-digit dialling (RDD) sampling design. To this end, the population was first stratified into 16 geographic strata corresponding to the 16 German states; see table 1 and figure 1. The 1500 smokers and 1000 non-smokers to be sampled were then divided amongst the strata using proportional allocation to the estimated size of the adult populations in each of the strata. This yielded quotas of smokers and non-smokers to be sampled in each stratum. Using the ADM-Master-Sample RDD frame, households were randomly called until the corresponding quota was met; this process was repeated independently for each stratum. A household was deemed eligible if it contained at least one eligible respondent (see above), but households with only cell phones were excluded. In households with multiple eligible respondents (this included non-smokers residing with smokers when the corresponding non-smoker quota was opened), the Next-Birthday method (Binson et al. (2000)) was used to select a single one. When the non-smoker quota of a given stratum was met, only households with one or more qualified smokers were deemed eligible. No substitution within household was allowed, except when it was known that the selected respondent would be absent for the entire fieldwork period.

The quotas were chosen to ensure representation proportional to the adult population of each state. However, fieldwork ended prematurely in Baden-Württemberg, Mecklenburg-Vorpommern and Niedersachsen, as these 3 states were the first to implement new tobacco control legislation on August 1, 2008. Consequently, these states have slightly smaller sample sizes than originally planned, and weight calculation has to be slightly adjusted (see section 2.2).

The ITC Germany wave 1 sample consists of 1515 adult smokers, and of 1059 adult non-smokers; for a total of 2574 respondents.

1.2 Wave 2

Out of the total of 2574 wave 1 respondents, 1002 smokers and 819 non-smokers were successfully re-contacted at wave 2; yielding a retention rate of 70.7% (66.1% for smokers and 77.3% for non-smokers). Like other ITC surveys, respondents lost to follow-up were suppose to be replenished by newly randomly selected ones, but lack of funding prevented this. Figure 2 shows the attrition of the ITC Germany samples over the first two waves.

Stratum #	German	English
1	Baden-Württemberg	Baden-Wuerttemberg
2	Bayern	Bavaria
3	Berlin	Berlin
4	Brandenburg	Brandenburg
5	Bremen	Bremen
6	Hamburg	Hamburg
7	Hessen	Hesse
8	Mecklenburg-Vorpommern	Mecklenburg-Vorpommern
9	Niedersachsen	Lower Saxony
10	Nordrhein-Westfalen	North Rhine-Westphalia
11	Rheinland-Pfalz	Rhineland-Palatinate
12	Saarland	Saarland
13	Sachsen	Saxony
14	Sachsen-Anhalt	Saxony-Anhalt
15	Schleswig-Holstein	Schleswig-Holstein
16	Thüringen	Thuringia

Table 1: Strata of the ITC Germany Survey RDD frame.

2 Weight construction

2.1 General comments about weight construction

As with most survey weights, the ITC Germany weights are constructed to correct and adjust for sample mis-representation caused by unequal sampling probabilities, frame error (i.e., under-coverage and multiplicity), and non-response as well as improving precision of estimates through the use of auxiliary information (e.g., smoking prevalences). In addition, conservative weight trimming was performed to prevent extreme weight variation arising from a few respondents having very large sampling weights. We briefly describe these key concepts of weight construction in this section, but refer the reader to Levy & Lemeshow (2008), chapter 16, for more detailed information.

At their base, sampling weights are defined as the inverse of selection probabilities, and thus adjust for sample mis-representation caused by unequal sampling probabilities. For example, a smoker residing alone has a probability of selection twice that of a smoker residing with another smoker.

Great efforts are made to create a complete/perfect sampling frame (i.e., a frame that would include all members of the target population, without duplicate and without any erroneous

Figure 1: Strata of the ITC Germany Survey RDD frame.

inclusions[1]). However, this is seldomly achieved and, consequently, some members of the target population are not part of the sampling frame (i.e., have a 0 probability of being selected). This is referred to as frame under-coverage, and can result in non-coverage bias. To reduce non-coverage bias in the ITC Germany Survey, post-stratification adjustments were performed on the sampling weights to ensure that, for each age/sex/region group, the totals of the sampling weights equal known benchmarks; see step 3 in section 2.2.1. Note that these benchmark figures are also referred to as calibration or target figures, and thus

[1]Erroneous inclusions refers to units that are not part of the target population, but included in the sampling frame.

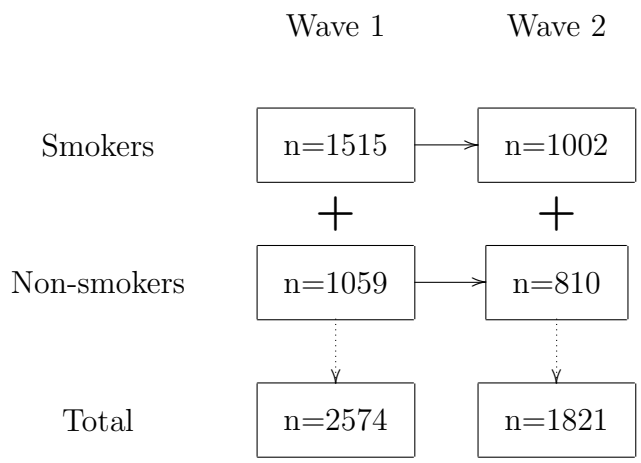

Figure 2: Attrition in the ITC Germany Survey.

the post-stratification adjustment is also referred to as weight calibration.

If non-respondents behave differently than respondents, then inference based solely on the sample of respondents will be biased unless adjustments are made. The greater the expected proportion of non-response, the greater this bias is likely to be. In the ITC Germany Survey, the post-stratification adjustments described in the above paragraph also adjust for non-coverage bias. It should be noted that if data are missing completely at random (MCAR, see Little & Rubin (2002)) within each age/sex/region group, then non-response bias will be completely eliminated. Realistically though, non-response bias is greatly reduced, but not eliminated in the ITC Germany Survey.

The distribution of sampling weights is often skewed to the right, echoing the fact that most populations are composed of many average/typical members and of few atypical ones. Average members have a fairly high probability of selection, and thus most sampling weights are fairly small. There are however few members of the population that have a much smaller probability of selection, and consequently have sampling weights that are quite large. These few large weights can be the source of high weight variation, which increases the variability of estimators. To correct for this, large weights are often trimmed in the weight construction process. This must be done with care and conservatively, as trimming can increase bias. There are various ways of trimming sampling weights. In the ITC Germany Survey, trimming was done by capping the number of adults (and thus the number of smokers) in each household at 4 (see step 1 in section 2.2.1). Capping is a fairly conservative weight trimming technique and, since it is done at the beginning of weight construction, helps minimize potentially biasing estimates.

It is well known from survey sampling theory that, in the vast majority of cases, the ratio estimator has much greater precision than the commonly used Horvitz-Thompson estimator. Heuristically, this is due to the fact that the ratio estimator utilizes auxiliary (i.e., additional) information in addition to the sampling weights, whereas the Horvitz-Thompson estimator does not. As mentioned above however, smoking prevalence figures were used to calibrate the ITC Germany sampling weights in order to reduce biases from frame errors and non-response. Our calibrating procedure yields (so-called) ratio weights, which enable all estimators to inherit the increased precision of the ratio estimator.

2.2 Wave 1 weights

Two sets of weights were computed at wave 1: cross-sectional weights for smokers and cross-sectional weights for non-smokers. Since no respondent can have both a smoker and non-smoker weight, both sets were combined into a single variable, labelled `aDE44919v`.

It should be noted that all ITC Germany survey weights were calibrated to smoking prevalence (see step 3 of section 2.2.1) and rescaled to have a mean equal to 1 (see step 4 of section 2.2.1). Consequently, these weights should not be used to estimate population totals (e.g., the total number of daily smokers). However, all of them can obviously be used to estimate population means and proportions/percentages, as well as in various statistical models (e.g., logistic and linear regressions).

2.2.1 Smoker weights

Computation of sampling weights for the 1515 smokers who completed the wave 1 survey proceeded as follows.

Step 1: Each respondent was first assigned an initial weight $w_i^{(1)}$, which can be viewed as an adjustment for the probability of selection within a given household while the non-smoker quota was opened and after it was closed. Formally, these $w_i^{(1)}$ weights are given by

$$w_i^{(1)} = \frac{\#\text{smokers}_i \times \#\text{adults}_i}{\hat{P}_{k(i)} \times \#\text{smokers}_i + (1 - \hat{P}_{k(i)}) \times \#\text{adults}_i} \ ,$$

where i stands for the ith respondent and $k(i)$ denotes the state/stratum to which that respondent belongs, $\#\text{smokers}_i$ is the number of adult smokers in the household, $\#\text{adults}_i$ is the number of adults (i.e., 18 years and over) in the household, and

$\hat{P}_{k(i)}$ is an estimate of the probability that the household of the ith respondent was called while the non-smoker quota was opened in state/stratum k ($k = 1, \ldots, 16$). Correspondingly, $1 - \hat{P}_{k(i)}$ is an estimate of the probability that the household of the ith respondent was called after the corresponding non-smoker quota was closed. Recall that $\#\text{adults}_i$ was capped at 4 to prevent large households from having undue influence on the weights; thus, $\#\text{smokers}_i \leq \#\text{adults}_i \leq 4$.

Computation of the $\hat{P}_{k(i)}$'s:

Let n_k be the total number of households in state k at which contact was made, and n_k^o be the number of households in state k at which contact was made while the non-smoker quota was open. Note that both n_k and n_k^o are readily available from call-logs. Then, $\hat{P}_{k(i)} = n_k^o/n_k$ is an estimator of the probability that the household of the ith respondent was called while the non-smoker quota was opened in state k, and this for all households in that state.

To produce weights that are less variable, the n_k's and n_k^o's were pooled across the states. However, as mentioned in section 1.1, fieldwork ended prematurely in Baden-Württemberg, Mecklenburg-Vorpommern and Niedersachsen. As a result, the $\hat{P}_{k(i)}$'s for these 3 states are fairly different from those of the other 13 states, and pooling over the 16 German states is not advisable. It was thus decided to pool the n_k's and n_k^o's into two groups: the 3 German states were fieldwork ended prematurely and the rest. This yielded $\hat{P}_{k(i)} = 0.391$ for those 3 states, and $\hat{P}_{k(i)} = 0.166$ for the remaining 13 states.

Step 2: A post-stratification adjustment was then performed to calibrate the $w_i^{(1)}$ weights to known proportions of the German adult population residing in each of the 16 states, as of July 20, 2007; see table A.2. For respondents residing in state k, this consisted in multiplying their $w_i^{(1)}$ weights by a factor f_k to produce adjusted $w_i^{(2)}$ weights. These $w_i^{(2)}$ weights are such that their sum over all respondents residing in state k divided by their sum over all respondents is equal to the known proportion in question. In other words,

$$w_i^{(2)} = w_i^{(1)} \times f_k \ ,$$

where

$$f_k = \frac{p_k}{\sum_{k \in F_k} w_i^{(1)} \Big/ \sum_{i=1}^{n} w_i^{(1)}} \ ,$$

and where $p_1, \ldots p_{16}$ are given in column 3 of table A.2 and F_k is the set of all respondents residing in state k. This was done to compensate for differential achieved sampling fractions from stratum to stratum.

Step 3: The $w_i^{(2)}$ weights were then calibrated to smoking prevalence by age/sex/region groups using the same post-stratification technique used in step 2. To this end, age was divided into 4 intervals (i.e., [18,25), [25,40), [40,55) and [55,100)); whereas the 16 states were grouped into 5 geographic regions: Eastern, Western, Southern, Northern and Middle (see table A.1). The resulting $w_i^{(3)}$ weights thus sum to the estimated number of adults smokers in each of the 40 age/sex/region cells of table A.3. Formally,

$$w_i^{(3)} = w_i^{(2)} \times \frac{c_\ell}{\sum_{i \in C_\ell} w_i^{(2)}} ,$$

where $c_1, \ldots, c_{40}$ are given in table A.3 and C_ℓ is the set of all respondents in cell ℓ.

The calibration figures of table A.3 were obtained by combining population estimates (as of Dec. 31, 2006) from Statistisches Bundesamt (i.e., the Federal Statistical Office of Germany) to prevalence estimates from the 2005 German Mikrozensus.

Step 4: To facilitate comparisons across multiple ITC countries, the $w_i^{(3)}$ weights were rescaled to have a mean equal to 1 or, equivalently, to sum to $n = 1515$ (the number of smokers who completed the wave 1 survey). This yielded the $w_i^{(4)}$ weights, which are formally defined as

$$w_i^{(4)} = w_i^{(3)} \times \mathrm{RF} ,$$

where RF is a rescaling factor and is given by

$$\mathrm{RF} = \frac{n}{\sum_{i=1}^{n} w_i^{(3)}} = \frac{1515}{\sum_{i=1}^{n} w_i^{(3)}} .$$

2.2.2 Non-smoker weights

Computation of sampling weights for the 1059 non-smokers who completed the wave 1 survey proceeded alike for the smokers weights; i.e.,

Step 1: Each respondent was first assigned an initial weight $w_i^{(1)}$, which can be viewed as an adjustment for the probability of selection within a given household while the non-smoker quota was opened. Using the same notation as in step 1 of section 2.2.1, these $w_i^{(1)}$ weights are formally given by

$$w_i^{(1)} = \frac{\#\mathrm{adults}_i}{\hat{P}_{k(i)}} ,$$

where $\hat{P}_{k(i)} = 0.391$ for Baden-Württemberg, Mecklenburg-Vorpommern and Niedersachsen, and $\hat{P}_{k(i)} = 0.166$ for the remaining 13 states.

Step 2: A post-stratification calibration was performed to compensate for differential achieved sampling fractions from state to state, same as step 2 of section 2.2.1.

Step 3: The weights were then calibrated to non-smoking prevalence by age/sex/region groups. This was done the same way as step 4 of section 2.2.1 with the exception that age was divided into 3 intervals (i.e., [18,40), [40,55) and [55,100)) instead of 4, and that the calibration figures for non-smokers were use instead of those of smokers. Hence, the resulting weights sum to the estimated number of adults non-smokers in each of the 30 age/sex/region cells of table A.3. These non-smoking calibration figures were obtained by combining the same two surveys as in step 4 of section 2.2.1, and by simply taking 1 minus the smoking prevalence figure as the corresponding non-smoking prevalence figure.

Step 4: The weights computed in step 3 above were rescaled to sum to sample size $n = 1059$, same as step 4 of section 2.2.1.

2.3 Wave 2 weights

Four sets of weights were computed for wave 2 data.

- **Waves 1–2 longitudinal weights for smokers** were computed for the 1002 smokers recruited at wave 1 that were retained and interviewed at wave 2. Note that 125 of these 1002 respondents had quit smoking at wave 2. Hence, if a respondent is recruited as a "smoker", he/she will always be considered as a "smoker" when computing his/her longitudinal weights and this regardless of his/her current smoking status.

- **Wave 2 cross-sectional weights for smokers** were computed for the 1037 respondents recruited at wave 1 that were smokers when interviewed at wave 2. These 1037 respondents consist of the 1002 smokers mentioned above, and of 35 respondents that were non-smokers at wave 1, but are now smoking. Hence, if a respondent was recruited as a "non-smoker" but became a smoker at wave 2, he/she will receive a cross-sectional smoker weight at wave 2.

- **Waves 1–2 longitudinal weights for non-smokers** were computed for the 819 non-smokers recruited at wave 1 that were retained and interviewed at wave 2. Note that 35 of these 819 respondents had started smoking at wave 2. As with smokers,

if a respondent is recruited as a "non-smoker", he/she will always be considered as a "non-smoker", when computing his/her longitudinal weights and this regardless of his/her current smoking status.

- **Wave 2 cross-sectional weights for non-smokers** were computed for the 784 non-smokers recruited at wave 1 that were still non-smokers when interviewed at wave 2. Hence, no wave 2 cross-sectional weights for non-smokers were computed for the 35 respondents that had started smoking at wave 2.

The two sets of longitudinal weights were constructed to adjust for attrition between waves 1 and 2, thus ensuring that the subset of respondents who completed both waves still represents the population at the time of wave 1 (i.e., Jul.–Nov. 2007). Hence, these wave 2 longitudinal weights were calibrated using wave 1 figures (i.e., table A.2, column 3, and table A.3). The two sets of cross-sectional weights were constructed for the same subset of respondents who completed both waves. However, new calibration figures (i.e., table A.2, column 4, and table A.4) were used to ensure that these respondents represent the population at the time of wave 2 (i.e., Jul.–Oct. 2009).

Since no respondent can have both a smoker and non-smoker longitudinal weights, both sets of longitudinal weights were combined into a single variable, labelled `bDE44921v`. Similarly, both sets of cross-sectional weights were combined into a single variable, labelled `bDE44919v`. As mentioned in section 2.2, all ITC Germany weights were calibrated to smoking prevalence by age/sex/region and rescaled to have a mean equal to 1. Consequently, these weights should not be used to estimate population totals (e.g., the total number of daily smokers).

2.3.1 Smoker weights

Starting with $w_i^{(0)}$, the wave 1 smoker weight for the ith respondent (computed in section 2.2.1), calculation of the 1002 **waves 1–2 longitudinal weights for smokers** proceeded as follows.

Step 1: The $w_i^{(0)}$ weights were first re-calibrated to the proportions of adults residing in each of the 16 German states, as of July 20, 2007; see column 3 of table A.2. This was done the same way as step 2 of section 2.2.1.

Step 2: The weights were then re-calibrated to smoking prevalence by age/sex/region groups. This was done the same way as step 3 of section 2.2.1 with the exception that the 40 cells of table A.3 were collapsed into 38 cells because of attrition; see note at foot of table A.3.

Step 3: Lastly, the weights were rescaled to sum to sample size $n = 1002$, same as 4 of section 2.2.1.

Calculation of the 1037 **wave 2 cross-sectional weights for smokers** proceeded as follows.

Step 1: Each respondent was assigned a starting weights $w_i^{(0)}$:

- if the ith respondent was a smoker at wave 1, $w_i^{(0)}$ was taken to be his/her wave 1 smoker weight (computed in section 2.2.1);
- if the ith respondent is one of the 35 respondents who were non-smokers at wave 1 but are now smoking at wave 2, $w_i^{(0)}$ was taken to be mean of the wave 1 smoker weights in the corresponding age/sex/state group.

Step 2: If the ith respondent is one of the 10 wave 1 smokers who moved to a new state between waves 1 and 2, his/her $w_i^{(0)}$ was replaced by the average of the wave 1 smoker weights of respondents in the same age/sex group living in his/her new state.

Step 3: The $w_i^{(0)}$ weights were calibrated to the proportions of adults residing in each of the 16 German states, as of Dec. 31, 2008; see column 4 of table A.2. This was done the same way as step 2 of section 2.2.1.

Step 4: The weights were then calibrated to smoking prevalence by age/sex/region groups. This was done the same way as step 3 of section 2.2.1, but using the updated figures given in table A.4 (with a total of 38 cells) instead of those in table A.3.

Step 5: Lastly, the weights were rescaled to sum to sample size $n = 1037$, same as 4 of section 2.2.1.

2.3.2 Non-smoker weights

Starting with $w_i^{(0)}$, the wave 1 non-smoker weight for the ith respondent (computed in section 2.2.2), calculation of the 819 **waves 1–2 longitudinal weights for non-smokers** proceeded as follows.

Step 1: The $w_i^{(0)}$ weights were first re-calibrated to the proportions of adults residing in each of the 16 German states, as of July 20, 2007; see see column 3 of table A.2. This was done the same way as step 2 of section 2.2.1.

Step 2: The weights were then re-calibrated to non-smoking prevalence by age/sex/region cells; see table A.3. This was done the same way as step 3 of section 2.2.2

Step 3: Lastly, the weights were rescaled to sum to sample size $n = 819$, same as 4 of section 2.2.1.

Starting with $w_i^{(0)}$, the wave 1 non-smoker weight for the ith respondent (computed in section 2.2.2), calculation of the 784 **wave 2 cross-sectional weights for non-smokers** proceeded as follows.

Step 1: If the ith respondent is one of the 3 non-smokers who moved to a new state between waves 1 and 2, his/her $w_i^{(0)}$ was replaced by the average of the wave 1 non-smoker weights of respondents in the same age/sex group living in his/her new state.

Step 2: The $w_i^{(0)}$ weights were calibrated to the proportions of adults residing in each of the 16 German states, as of Dec. 31, 2008; see see column 4 of table A.2. This was done the same way as step 2 of section 2.2.1.

Step 3: The weights were then calibrated to non-smoking prevalence by age/sex/region groups. This was done the same way as step 3 of section 2.2.2, but using the updated figures given in table A.4 (with a total of 16 cells) instead of those in table A.3.

Step 4: Lastly, the weights were rescaled to sum to sample size $n = 784$, same as 4 of section 2.2.1.

Acknowledgements

Core funding for the ITC Project is provided by the U.S. National Cancer Institute (NCI) to the Roswell Park TTURC (P50 CA111236 & P01 CA138389), the Canadian Institutes of Health Research (CIHR; Grant #79551), and by the Ontario Institute for Cancer Research. Major funding for the ITC Germany Survey provided by Deutsches Krebsforschungszentrum (DKFZ; German Cancer Research Center), Bundesministerium für Gesundheit (German Federal Ministry of Health), and Dieter Mennekes-Umweltstiftung.

References

Binson, D., Canchola, J. A. & Catania, J. A. (2000), 'Random selection in a national telephone survey: a comparison of the kish, next-birthday and last-birthday methods', *Journal of Official Statistics* **16**, 53–60.

Levy, P. S. & Lemeshow, S. (2008), *Sampling of Populations: Methods and Applications*, 4 edn, John Wiley & Sons, Hoboken, NJ.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, John Wiley & Sons, Hoboken, NJ.

Statistisches Bundesamt (2005), 'Mikrozensus'.

Statistisches Bundesamt (2007), 'Population estimates/projections as of Jul. 20, 2007', https://www-genesis.destatis.de/genesis/online/.

Statistisches Bundesamt (2008), 'Population estimates/projections as of Dec. 31, 2008', https://www-genesis.destatis.de/genesis/online/ (accessed Dec. 9, 2009).

Appendix: Benchmark/calibration figures

The estimated number of smokers and non-smokers given in tables A.3 and A.4 were obtained by combining population estimates from Statistisches Bundesamt (i.e., the Federal Statistical Office of Germany) to prevalence estimates from the 2005 German Mikrozensus. To estimate the number of smokers, population estimate for a given age/sex/region combination were simply multiplied by the smoking prevalence for the same age/sex/region combination. Estimation of the number of non-smokers proceeded the same way, except that population estimates were multiplied by 1 minus the smoking prevalence of the corresponding age/sex/region combination. Some age/sex/region combinations were collapsed because they contained too few respondents. Population estimate as of Dec. 31, 2006 were used for table A.3, whereas figures as of Jul. 20, 2007 were used for table A.4.

Region	**States**
Southern	Bayern & Baden-Württemberg
Eastern	Sachsen, Sachsen-Anhalt, Brandenburg & Berlin
Northern	Niedersachsen, Bremen, Hamburg, Schleswig-Holstein & Mecklenburg-Vorpommern
Western	Nordrhein-Westfalen
Middle/Central	Saarland, Rheinland-Pfalz, Hessen & Thüringen

Table A.1: Grouping of the 16 German states into 5 geographic regions.

APPENDIX

Stratum #	State	Proportion (%) as of 20/07/2007	Proportion (%) as of 31/12/2008
1	Baden-Württemberg	12.99	12.90
2	Bayern	15.08	15.09
3	Berlin	4.13	4.30
4	Brandenburg	3.09	3.21
5	Bremen	0.81	0.82
6	Hamburg	2.13	2.20
7	Hessen	7.37	7.36
8	Mecklenburg-Vorpommern	2.08	2.12
9	Niedersachsen	9.70	9.53
10	Nordrhein-Westfalen	21.91	21.61
11	Rheinland-Pfalz	4.92	4.87
12	Saarland	1.28	1.27
13	Sachsen	5.21	5.36
14	Sachsen-Anhalt	3.02	3.05
15	Schleswig-Holstein	3.43	3.41
16	Thüringen	2.85	2.90
Total		100	100

Table A.2: German adult (i.e., 18+) population by state.

Region	Sex	Age	#smokers	#non-smokers
Eastern	male	[18,25)	280425	316133 [a]
Eastern	male	[25,40)	581161	698980 [a]
Eastern	male	[40,55)	661753	918429
Eastern	male	[55,100)	348737	1490437
Middle	male	[18,25)	240740	318539 [b]
Middle	male	[25,40)	546541	766218 [b]
Middle	male	[40,55)	629670	1015340
Middle	male	[55,100)	334921	1583980
Northern	male	[18,25)	287102 [†]	327188 [c]
Northern	male	[25,40)	635401 [†]	851210 [c]
Northern	male	[40,55)	757699	1008132
Northern	male	[55,100)	445309	1659872
Southern	male	[18,25)	377613	583386 [d]
Southern	male	[25,40)	901307	1455372 [d]
Southern	male	[40,55)	938604	1818207
Southern	male	[55,100)	533653	2585341
Western	male	[18,25)	285777	450203 [e]
Western	male	[25,40)	755870	1002672 [e]
Western	male	[40,55)	887396	1240917
Western	male	[55,100)	513118	1962411
Eastern	female	[18,25)	209973	340845 [f]
Eastern	female	[25,40)	398038	760412 [f]
Eastern	female	[40,55)	457729	1053076
Eastern	female	[55,100)	206619	2170139
Middle	female	[18,25)	175580	363418 [g]
Middle	female	[25,40)	374051	896520 [g]
Middle	female	[40,55)	482371	1109883
Middle	female	[55,100)	228905	2130601
Northern	female	[18,25)	215282 [‡]	377489 [h]
Northern	female	[25,40)	478479 [‡]	951769 [h]
Northern	female	[40,55)	581715	1124856
Northern	female	[55,100)	295821	2305293
Southern	female	[18,25)	301824	641775 [i]
Southern	female	[25,40)	624395	1691582 [i]
Southern	female	[40,55)	723287	1945173
Southern	female	[55,100)	349122	3475304
Western	female	[18,25)	231418	485150 [j]
Western	female	[25,40)	593893	1143273 [j]
Western	female	[40,55)	734222	1356639
Western	female	[55,100)	368662	2722449

Cells sharing the same letter (a, b, etc.) were collapsed, when computing the non-smoker weights. Cells sharing the same symbol († & ‡) were collapsed, when computing waves 1–2 longitudinal smoker weights.

Table A.3: Estimated # of smokers and non-smokers, per age/sex/region, used for calibration of wave 1 weights and waves 1–2 longitudinal weights.

APPENDIX

Region	Sex	Age	#smokers	#non-smokers
Eastern	male	[18,25)	268731	302951 [a]
Eastern	male	[25,40)	562687	676150 [a]
Eastern	male	[40,55)	658236	912925
Eastern	male	[55,100)	361606	1552650
Middle	male	[18,25)	240153	319237 [b]
Middle	male	[25,40)	517094	722333 [b]
Middle	male	[40,55)	636510	1029416
Middle	male	[55,100)	344681	1636759
Northern	male	[18,25)	289379 [k]	332347 [c]
Northern	male	[25,40)	597276 [k]	798440 [c]
Northern	male	[40,55)	779247	1039720
Northern	male	[55,100)	456820	1712612
Southern	male	[18,25)	386358	597351 [d]
Southern	male	[25,40)	857963	1379658 [d]
Southern	male	[40,55)	969680	1883208
Southern	male	[55,100)	547326	2674380
Western	male	[18,25)	291539	459421 [e]
Western	male	[25,40)	709970	940278 [e]
Western	male	[40,55)	909836	1274797
Western	male	[55,100)	523573	2011986
Eastern	female	[18,25)	203544	330183 [f]
Eastern	female	[25,40)	384547	738564 [f]
Eastern	female	[40,55)	453562	1042267
Eastern	female	[55,100)	212276	2214641
Middle	female	[18,25)	174598	362913 [g]
Middle	female	[25,40)	354827	850557 [g]
Middle	female	[40,55)	487714	1124182
Middle	female	[55,100)	234310	2168286
Northern	female	[18,25)	215695 [l]	379716 [h]
Northern	female	[25,40)	453052 [l]	901069 [h]
Northern	female	[40,55)	597482	1157674
Northern	female	[55,100)	300347	2342272
Southern	female	[18,25)	304864	649561 [i]
Southern	female	[25,40)	598756	1616631 [i]
Southern	female	[40,55)	745790	2008950
Southern	female	[55,100)	357180	3544816
Western	female	[18,25)	234564	491923 [j]
Western	female	[25,40)	561064	1079499 [j]
Western	female	[40,55)	751971	1390259
Western	female	[55,100)	374899	2758550

Cells sharing the same letter (a, b, etc.) were collapsed, when computing the wave 2 weights.

Table A.4: Estimated # of smokers and non-smokers, per age/sex/region, used for calibration of wave 2 cross-sectional weights.