pairfam
Panel Analysis of Intimate Relationships and Family Dynamics

New weights for the pairfam anchor data

Martin Wetzel, Nina Schumann, Claudia Schmiedeberg

May 2021

Funded as long-term project by the German Research Foundation (DFG)

Cite as:
Wetzel, Martin, Nina Schumann, and Claudia Schmiedeberg (2021): New weights for the pairfam anchor data. pairfam Technical Paper No. 17. https://doi.org/10.5282/ubm/epub.91999

pairfam Technical Paper
17

1 Introduction

The pairfam group in collaboration with the GESIS Team on Survey Statistics[1] revised the weighting procedures of the pairfam data with Release 12.0, now providing new, easy-to-use weights for all anchor data sets[2]. The revised procedures employ a more transparent approach that is better-suited to handling the longitudinal data structure, selective participation, and the different sampling frames than the previous weighting approach.

Compared to the former approach, the new weights have been improved in several ways:

- The new calibration procedure that adjusts to information from the Mikrozensus for each wave ensures that the data more closely represent the (changing) population of interest in central characteristics and size.
- The new weighting procedure addresses cohort size relative to each other, which leads to more adequate representations of cohort sizes in the weighted sample.
- The variances of the calibrated weights now remain relatively small over observations.
- The choice of variables and the impact of each variable on the weights are now transparent.
- Using the new weights is now easier than ever, as it requires no further preparations by the user (such as multiplying different types of weights).

This Technical Paper provides details regarding the calculation and characteristics of the new pairfam weights to guarantee maximum transparency for users. Section 2 outlines how the weights were calculated, followed by a presentation of several basic characteristics of each weight in Section 3, in particular the distribution of the weights and development over the panel. Section 4 discusses the impact of the reference characteristics on each of the weights, indicating which subgroups have been over- or underrepresented in the sample. In Section 5, the variance of the weights is analysed. Section 6 presents comparisons for two (un)weighted variables with the Mikrozensus distributions, and Section 7 gives recommendations for use.

Further examples of using the pairfam weights are available in the Quick Start do-file "Weighting", and a summary of the weighting approach is included in Section 4.6 of the pairfam Data Manual. General information regarding different weighting approaches can be found, for instance, in Lavallée and Beaumont (2015). Methodological details about the new weighting approach applied to the pairfam data are given in Lundström and Särndal (1999).

[1] In particular, we would like to thank Dr. Bruch, Dr. Felderer, and Dr. Sand of GESIS for their consulting and support in implementing the new weighting strategy. For all remaining issues the pairfam team is fully responsible.
[2] For more information about the previous weighting approach, see the Data Manual of former pairfam releases (11.0 and earlier).

2 Weighting procedures

Two types of weights are provided as part of the anchor data sets: design weights (*d*weight* with * representing different samples) and calibrated design weights (*cd*weight*). Design weights aim to correct for disproportionate sampling across cohorts and the combination of multiple selection frames (pairfam base sample, DemoDiff sample, and wave 11 refreshment). Calibrated design weights aim to correct for survey non-response (including longitudinal panel drop-out). Note that instead of specific calibration weights, pairfam provides ready-to-use combinations of calibration and design weights, calibrated design weights, as using design weights is a precondition for the calibration. Users therefore do not need to combine these further.

In the pairfam survey, three samples have been drawn from three different frames:

1) pairfam base frame: Representative of the German population from three age cohorts (1971-73, 1981-83, 1991-93),
2) DemoDiff frame: Representative of the population living in Eastern Germany from two age cohorts (1971-73, 1981-83),
3) pairfam refreshment frame: Representative of the German population from three age cohorts (1981-83, 1991-93, 2001-03).

Accordingly, inclusion probabilities must be addressed for analyses wishing to include all three samples into one data set. This can be done by identifying the different frames and applying a composite estimator in which weights of multiple frames are combined in the ratio of their respective net sample sizes (Brick, Cervantes, Lee, & Norman, 2011; Lohr & Rao, 2000; Sand, 2018). For this purpose, four versions are provided for each weight, computed in the same way but based on different samples:

1) *dweight/cdweight:* To be used if analyzing the pairfam base sample only
2) *d1weight/cd1weight:* To be used when analyzing the pairfam base sample and DemoDiff sample together
3) *d2weight/cd2weight:* To be used when analyzing pairfam base, DemoDiff and refreshment samples together
4) *d3weight/cd3weight:* To be used when analyzing the refreshment sample only.

Note that weights are only provided for the sampled primary (so-called “anchor respondents”) of the four cohorts, while no weights are available for step-up respondents or secondary respondents (i.e., partners, children, and parents).

2.1 Design weights

Due to pairfam's cohort study design, the design weight is the factor by which each of the birth cohorts are under- or overrepresented in the gross sample as compared to the target population. In their simplest form, the design weights represent the cohort-specific inverse probability to be included in the gross sample based on the size of the target population. Design weights *w* are calculated and standardized step-wise for each cohort *i* (and at each wave t) as follows, so that $\bar{w}_t = 1$ and $\sum w_t = N_{T,NS}$:

$$w_i = \frac{N_{i,P} N_{T,NS}}{N_{i,GS} \sum_{j=1}^{3} \left(\frac{N_{j,P} N_{j,NS}}{N_{j,GS}} \right)}$$

with $N_{i,P}$ the size of cohort *i* in the population, $N_{T,P}$ the total size of all cohorts in the population, $N_{i,GS}$ the size of cohort *i* in the gross sample, $N_{T,GS}$ the total size of all cohorts in the gross sample, $N_{i,NS}$ the size of cohort *i* in the net sample (of their first appearance), and $N_{T,NS}$ the total size of all cohorts in the net sample.

Figure 1 Gross samples of pairfam relative to population size illustrates the challenges for the design weights by presenting gross sample and population sizes. In wave 1, the target population sizes of the three birth cohorts 1991-1993, 1981-1983, and 1971-1973 in Germany in 2008 were 2.54 million, 2.99 million, and 3.12 million, respectively. According to the pairfam methods report (Suckow & Schneekloth, 2009), the gross respective sample sizes are 9,648, 16,810, and 15,616. Although the population sizes for each cohort varied only slightly, the youngest cohort was targeted with a considerably smaller gross sample due to the fact that response rates were expected to be higher than in the older cohorts (Brüderl, Schmiedeberg, et al., 2021). This approach led to relatively similar sample sizes over the three cohorts in the pairfam base study, but the design weight factors vary significantly (1.225, 0.828, and 0.930 in wave 1). An additional sample of people living in the Eastern Germany (DemoDiff) was integrated post-hoc into the pairfam study in wave 3. If both samples are analyzed together without weighting, residents of Eastern Germany would be over-represented and the youngest cohort under-represented, as this cohort was not included in the DemoDiff sample.

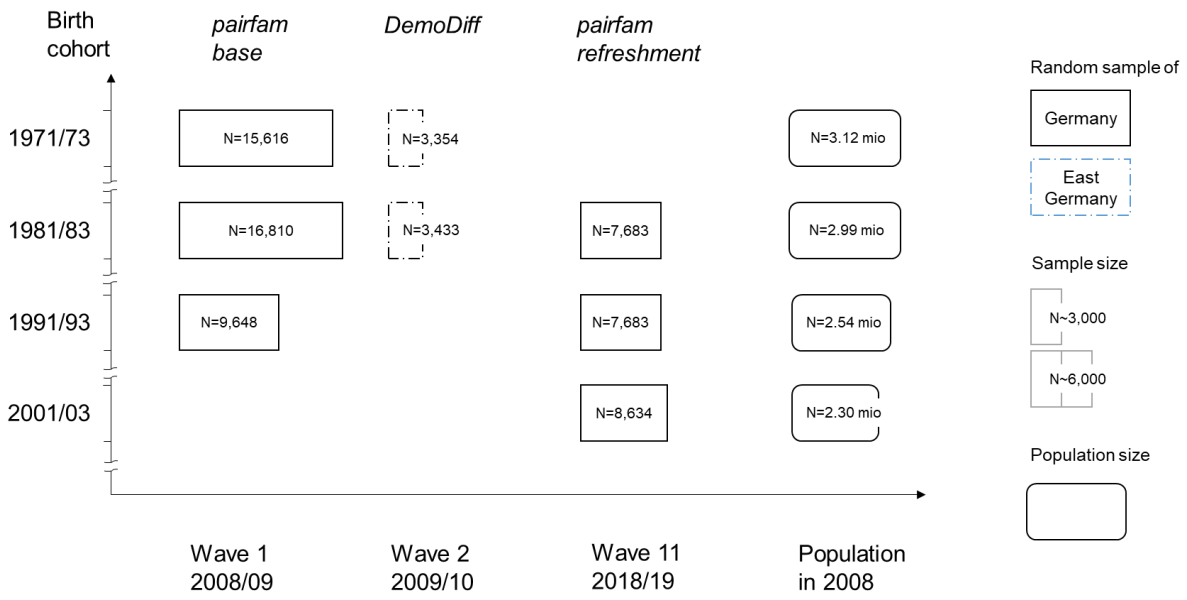

Figure 1 Gross samples of pairfam relative to population size

As part of the wave 11 refreshment sample, no new cases were drawn for the oldest pairfam/DemoDiff cohort (1970-73), and a new, younger cohort (2000-03) was integrated into the sample. Hence, an unweighted analysis of the complete wave 11 data would result in the oldest cohort being under- and the youngest over-represented (for an example see Figure 9).

In sum, the design weights have been defined to achieve the following:

- Correct for different inclusion probabilities of each cohort, i.e. under- or over-representation in the gross sample as compared to the target population,
- Integrate the DemoDiff sample into the pairfam base sample by addressing the shares of respondents living in Eastern and Western Germany,
- Address the issues arising from drawing a refreshment sample with both distinct and overlapping sampling frames.

2.2 Calibrated design weights

Calibration weights provide factors to adjust the observed data in each subpopulation to characteristics of the target population (Deville, Särndal, & Sautory, 1993; Lundström & Särndal, 1999). This approach meets three goals: First, it ensures that the weighted data more closely represent the population of interest in central characteristics and size. Second, selective non-response can be managed by assigning higher analysis weights to observations with characteristics of higher selectivity, tackling both cross-sectional survey participation bias and longitudinal panel attrition bias. Third, a correction of cohort-specific non-response can be integrated so that weighted data represents cohort sizes relative to their counterparts in the target population. As calibration weights can be applied only in combination with design

weights, the pairfam group provides combined calibration and design weights as *calibrated design weights* instead of separate calibration weights.

To achieve calibration in pairfam, an iterative proportional fitting (IPF) approach was applied using the *ipfraking* package (Kolenikov, 2014, 2019) for Stata to successively identify higher or lower weights until an optimal adjustment to the reference data was achieved.[3] Weights were computed cross-sectionally, meaning that identical IPF procedures were applied for each wave separately with data from the Mikrozensus of the respective year as a reference.

It is important to note that weighting reduces bias in a variable only if the reference characteristics are correlated both with the variable of interest and with non-response (Gabler, 2004, p. 141). Therefore, the goal here was to select reference characteristics for calibration that are likely to be associated with pairfam research topics and with non-response processes. As the calibration was based on data from the Mikrozensus of the respective years, the choice of variables was limited; only sociodemographic characteristics for which a reference distribution was available in all years of the pairfam study period (i.e., 2008-2020) could be considered. The following reference characteristics from the Mikrozensus were used:

- Gender: male, female
- Federal state (Bundesland): 14 categories[4]
- Education level: none or primary education (Hauptschule), lower secondary (Realschule), higher secondary (Abitur), still in education
- Migration background: none, first generation, second generation
- Settlement structure: 8 categories[5]
- Family status: single, married, widowed/divorced[*]
- Child(ren) living in household: none, one, two or more[*]

The pairfam data contain very few (<1%) missing values in the variables used for calibration. Single imputation was employed for these cases, meaning the missing values were replaced with the modus value of the respective variable.

[3] The replication files are provided as supplementary files of this Technical Paper.
[4] Germany has 16 federal states. To avoid small case numbers in cells, Saarland has been combined with Rheinland-Pfalz and Bremen with Hamburg.
[5] The original "753 systematic" of the BIK has 10 categories. To avoid small case numbers in cells, <2,000 and 2,000 to <5,000, as well as 50,000 to <100,000 rural and urban were collapsed into a joined category.
[*] These variables were not used for respondents under the age of 21, as the number of observations was not sufficient, leading to empty cells for particular characteristics in the population (Mikrozensus). Hence, the variables were not used for cohort 1 (born 1991-93) before wave 7 (2014), and never for cohort 4 (born 2001-03).

Calibration was stratified by cohort, but conducted in a joint model: Although the reference characteristics for each cohort are applied as the weighting target, the ratio of cohort sizes represents their ratio in the target population. Calibrated design weights were standardized at each wave t using the net sample size $N_{T,NS}$ so that the mean of all weights at t $\overline{w}_t = 1$ and the sum of all weights at t represent the sample size ($\sum w_t = N_{T,NS}$).

3 Basic characteristics

3.1 Design weights

As design weights do not account for individual characteristics, they are constant within sample frames; in the case of pairfam, within cohorts (2001-2003, 1991-1993, 1981-1983, 1971-1973) and samples (pairfam base, DemoDiff, refreshment). However, they vary slightly across waves as they are standardized to the net sample size. For instance, as shown in Figure 2 for the pairfam base sample, the design weight *dweight* differs by cohort and wave. If the sample also includes DemoDiff cases (*d1weight*), they differ additionally by region (Eastern vs. Western Germany). For the full sample, including refreshment cases (*d2weight*), they differ by time point of sample selection (base or refreshment), cohort, region, and wave.

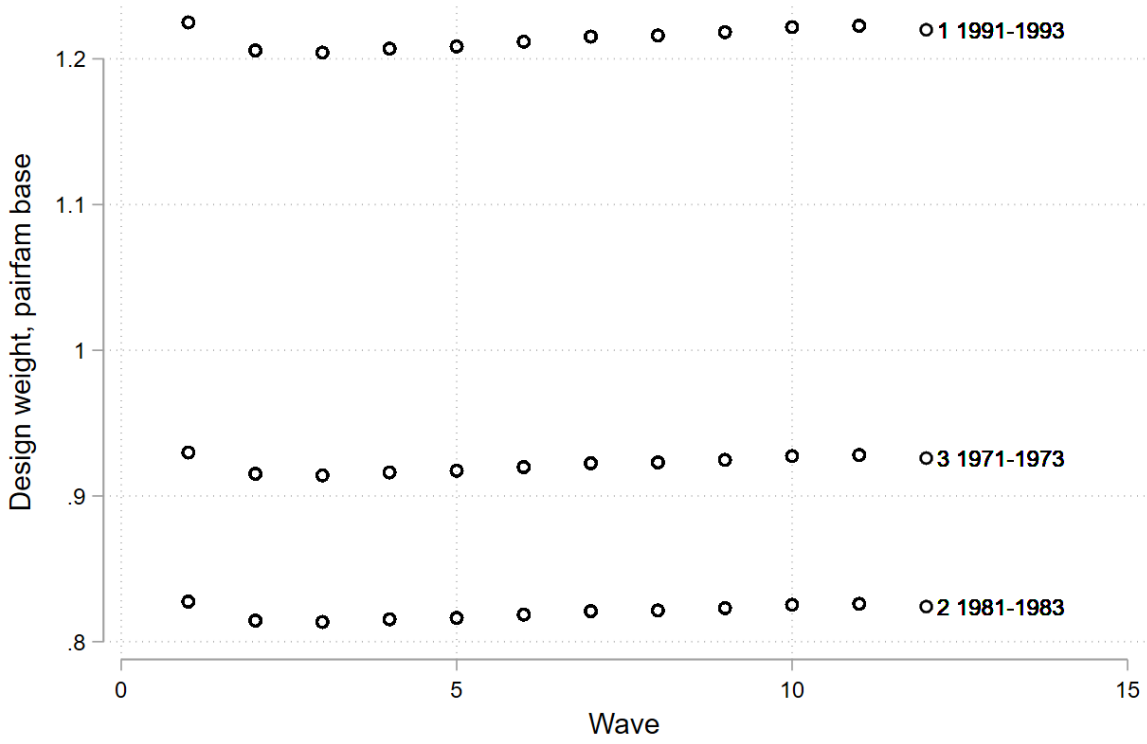

Figure 2: Distribution of the design weight dweight

3.2 Calibrated design weights

The distribution of the calibrated design weights is more granular, as each respondent is assigned a specific weight according to her/his individual characteristics. The calibrated design weights are standardized step-wise for each cohort (and at each wave) so that $\bar{w}_t = 1$ and $\sum w = N_{T,NS}$. As is typical for weights, values are near 1 for the majority of respondents, whereas a much smaller group of respondents (i.e., those with characteristics implying high non-response propensities) have larger weights.

Figure 3 shows median values and the distribution of the calibrated design weight for the pairfam base sample *cdweight* over time. Due to the standardization approach, the mean level of the weight is 1 in each wave. The displayed medians show over the whole period values <1, indicating a left-steep distribution. In wave 1, the standard deviation is 0.42 (12,402 participants). Up to wave 11, case numbers decline to 3,808 participants due to panel attrition (see also Brüderl, Schmiedeberg, et al., 2021 and Müller & Castiglioni, 2015, 2020). While the mean level of the weights remains at 1 due to the standardization, the standard deviation increases to 0.65 in wave 11, indicating that is more difficult and thus less efficient to reproduce the reference characteristics by using a weighted sample over time. This can be also seen in the increase in both number and size of extreme cases: Explorative analysis of extreme cases showed that some cross-combinations receive very high weights (e.g., childless respondents of the oldest cohort with low educational levels, respondents of the second cohort with high educational levels and migration background). All values are provided untrimmed.

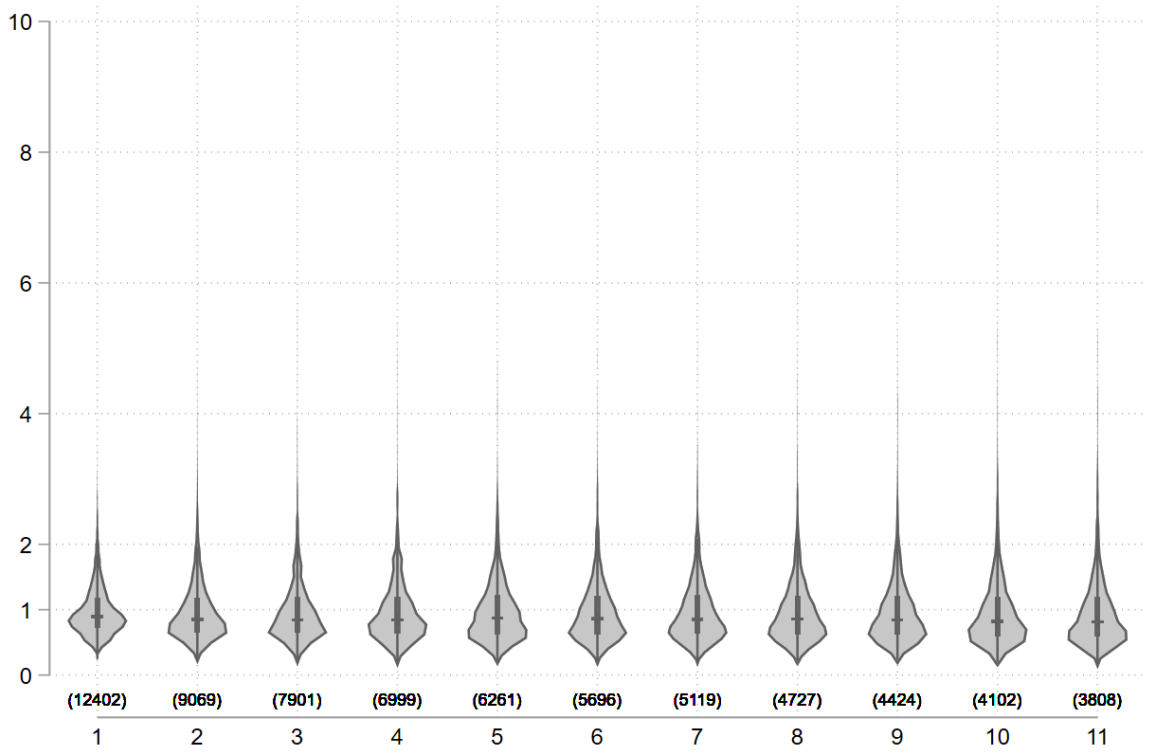

Figure 3: Violin plot showing the distribution of cdweight over waves

The distributions of the other calibrated design weights *cd1weight*, *cd2weight,* and *cd3weight* exhibit similar patterns across the panel (see Figure 11 in the Appendix).

Finally, to illustrate the impact of the weight *cd2weight* for cohort sizes, Figure 4 shows unweighted and weighted distributions for all cohorts compared to information from the Mikrozensus on the target population. Comparing the unweighted with the weighted distribution demonstrates the extent of over- and under-representation of each of the cohorts included in the pairfam study. Without weighting, the cohorts contribute partially too much/too little as compared to their “true” share in the German population. As the cohort variable is an essential part of the calibration procedure, the results from *cd*weight* are quite representative of the target population distribution.

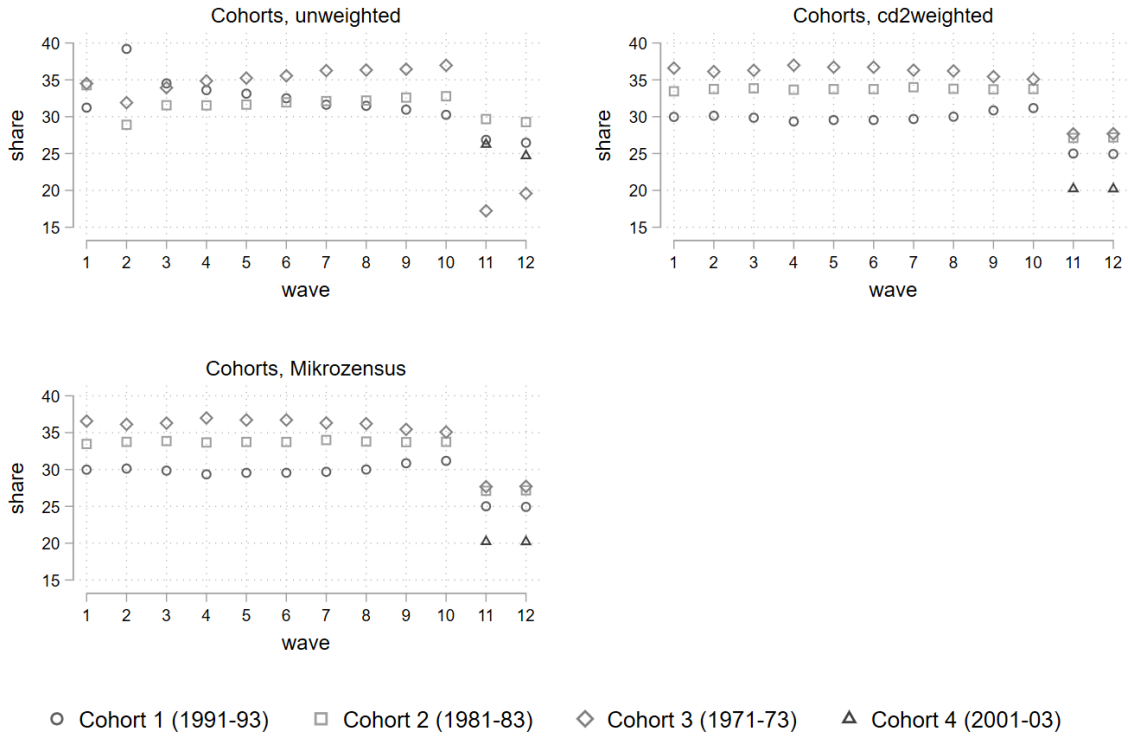

Figure 4: Cohort distribution (unweighted and cd2weighted) over panel waves

4 Contribution of reference characteristics on weight size

The following illustrates how the calibration variables contribute to the size of the weights. Unlike regression analysis, which could be presented using coefficient plots, the raking procedure applied to compute calibration weights does not natively produce an output indicating the contribution of each calibration variable. Therefore, mean values for each of the weights were calculated for different subsamples, differentiated by reference characteristics. Figure 5 displays these means for *cdweight*, *cd1weight*, and *cd2weight* in wave 1. Note that although the values are presented in one graph, mean levels were estimated for each characteristic (e.g., gender, migration background) separately.

Due to the standardization of the calibrated design weights to the mean of 1, changes in the contribution of the variables can be easily interpreted. Results for *cd1weight* and *cd2weight* start to differ with the refreshment sample (in wave 11), and *cd3weight* cannot be displayed because wave 1 does not include any observations from the refreshment sample.

As expected, patterns are similar for all three weights. Subgroup-specific differences in mean weights indicate that men are slightly unrepresented in the sample, leading to slightly larger weights for men than for women. The same holds for respondents with migration backgrounds, who were assigned slightly larger weights than respondents without migration backgrounds. A

rather large difference is found for educational level, as individuals who left school without a degree or only with primary education are underrepresented and accordingly receive larger weights. Similarly, respondents without children receive larger weights, as that they are slightly underrepresented in the pairfam samples. The most pronounced difference between *cdweight* and *cd1weight* is evident between the federal states, as *cd1weight* integrates the DemoDiff subsample of respondents living in Eastern Germany. These respondents are assigned lower weights when *cd1weight* is used to avoid over-representation.

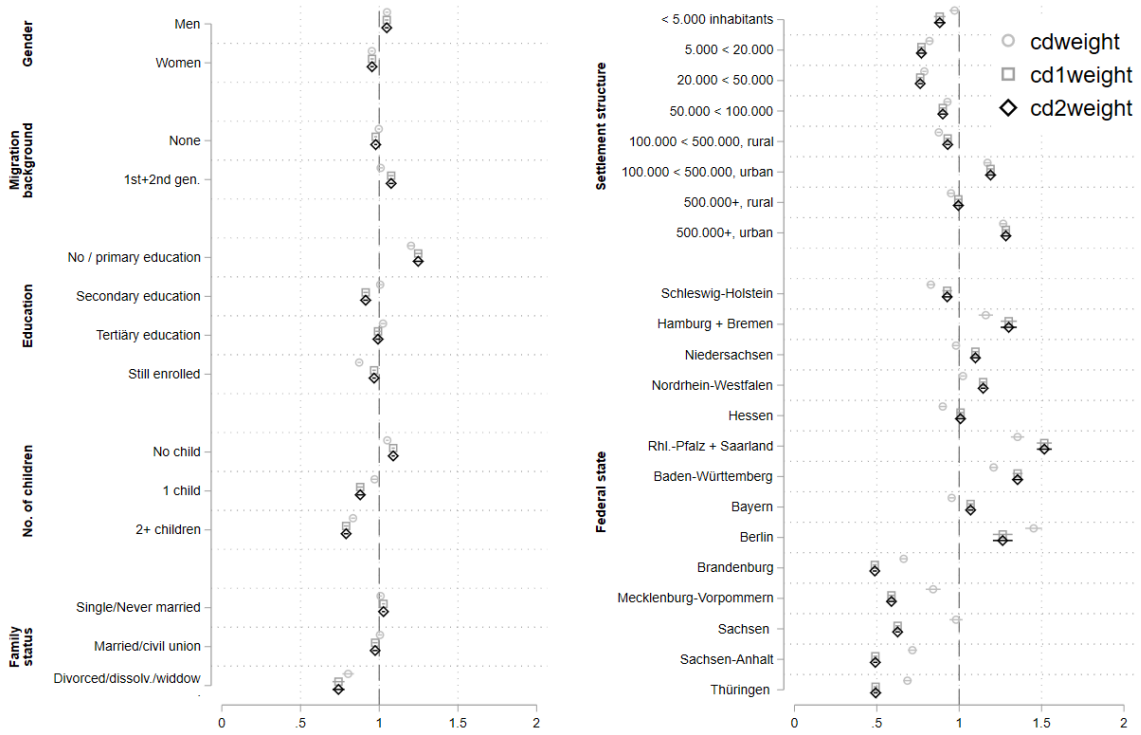

Figure 5: Means of calibrated design weights for wave 1

Figure 6 shows the mean levels of for all calibrated design weights *cd*weight* for wave 11 (also including *cd3weight*). The general patterns for the re-interviewed samples (considered by *cdweight*, *cd1weight*, and *cd2weight*) confirm the existing panel attrition literature: Men, individuals with migration backgrounds, and those with low educational levels have larger weights, as they drop out more often. The same pattern is found for adults currently enrolled in school. As expected for a family survey, individuals with children are more likely to participate and therefore receive smaller weights. Finally, respondents living in larger and more urban settlements are more difficult to re-interview, which is the case in particular for the federal state of Berlin. Calibrated design weights for the refreshment sample only (*cd3weight*) in wave 11 correct only for the initial participation bias at the first observation.

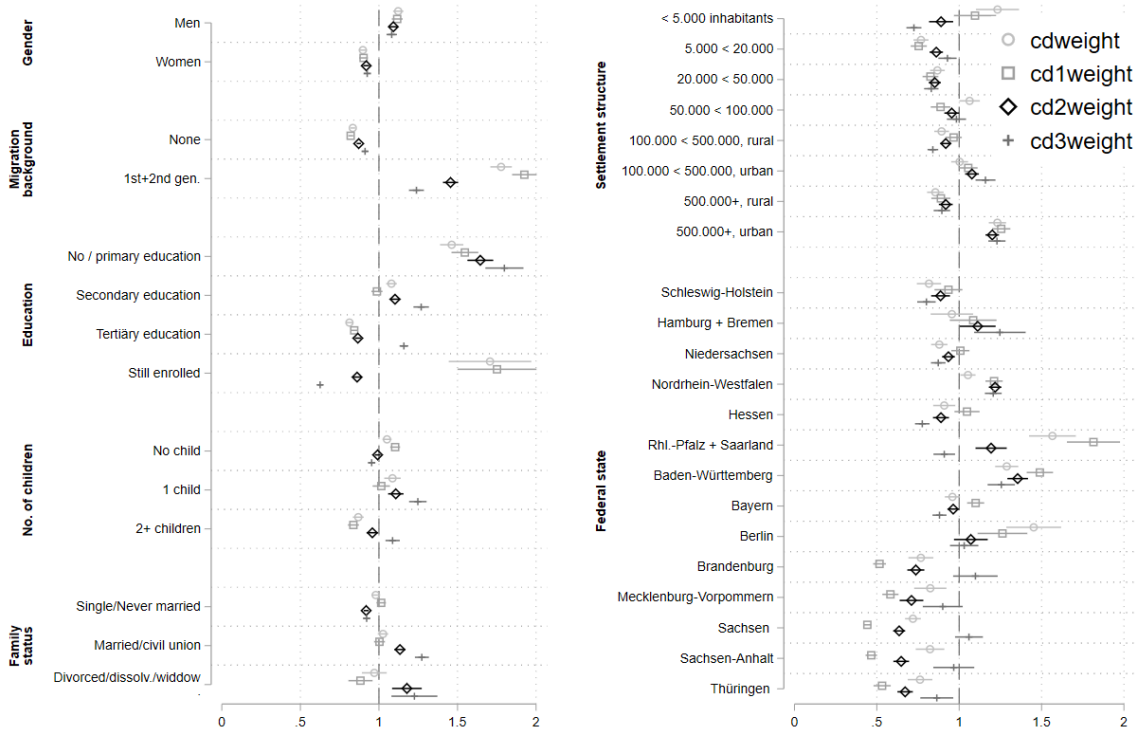

Figure 6: Means of calibrated design weights for wave 11

Calibration is conducted for each cohort separately based also on the design weights, which also differ by cohort (see p. 4). Accordingly, cohorts differ in calibrated design weight means, with smallest mean levels at wave 1 for the cohort born 1991-93 (0.96) and largest for the cohort born 1971-73 (1.06). Figure 7 displays the mean values of the weighting variable *cd2weight* for all cohorts for wave 1. Note that men of the oldest cohort in particular had a lower probability to participate, therefore receiving larger weights. Also, individuals of the oldest cohort who were (still) enrolled in education are underrepresented in the pairfam study as compared to the general German population, leading to larger weights for those respondents. In cohort 1, the number of children in the household and relationship status were not addressed in the weighting procedure up to wave 7 (analog for cohort 4 in wave 11) due to the low number of cases in the target population. Due to the DemoDiff sample with additional cases from Eastern Germany, weights for the Eastern federal states (Brandenburg, Thüringen) are much smaller.

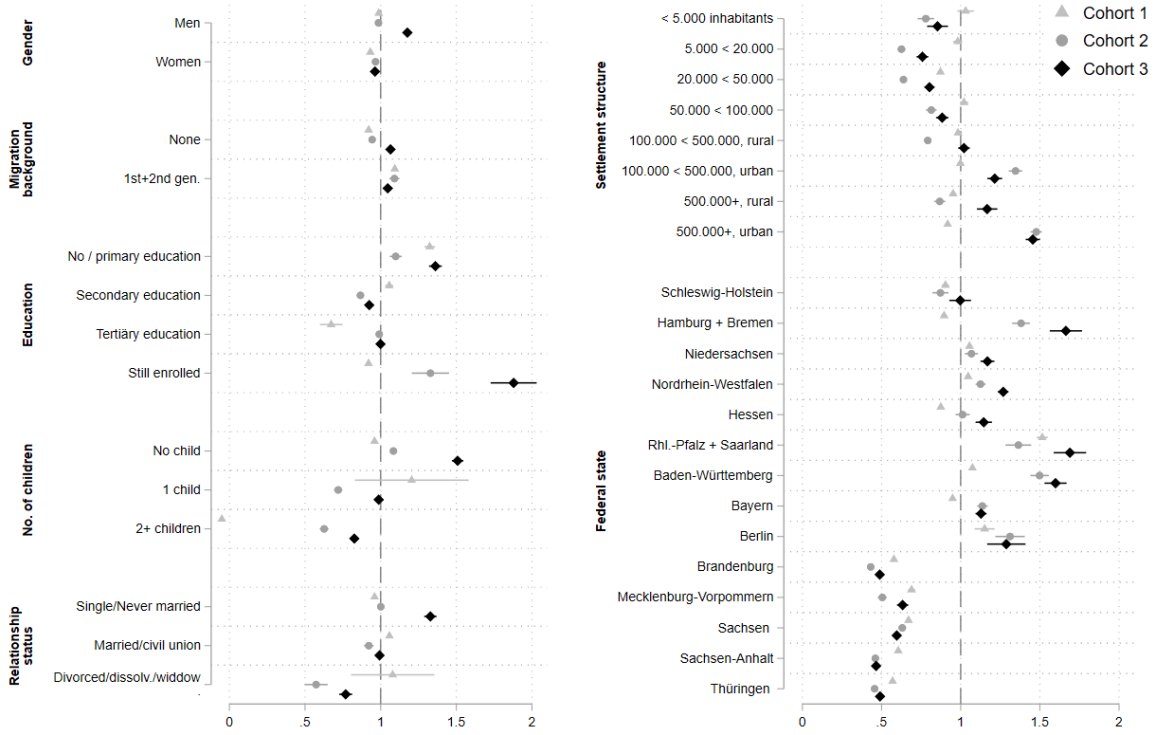

Figure 7: Means of cd2weight over cohorts for wave 1

5 Contribution of reference characteristics to variance in weights

Additional analyses show how the considered reference characteristics contribute to the variance in the weights. Weights with low variability are preferable, as low variability is associated with higher precision in weighted analyses (e.g., Larsen, Qing, Zhou, & Foulkes, 2011). A simple analysis of variance (ANOVA) is used here to identify parts of the variance of the *cd2weight* variable depending on the reference variable used for generating the weights. Results are displayed in Figure 8. Note that the overall level of variance is considerably higher in wave 11 than in wave 1, illustrated with the larger chart size. Results indicate that the variance at wave 1 is driven mainly by level of education (*school*), settlement structure (*bik*), federal state (*bula*), number of children (*kids*), and to a lesser extent by relationship status (*famstat*) and migration status (*migrat*), and to almost no extent by gender (*sex*). The design weight (*d2weight*) also contributes to the variance of *cd2weight*.

Compared to wave 1, variance in wave 11 is driven more strongly by migration background (*migrat*) and less by regional differences (*bula, bik*), number of children (*kids*), and family status (*famstat*). In both waves, a considerable part of the variance, termed “residuals”, is due to cross-combinations of characteristics (i.e., interaction effects) in addition to the main effects of the variables displayed in Figure 8.

The selection of the reference characteristics is a trade-off situation between increasing variance in the weights and reducing the potential selection bias in the sample. This analysis shows that the patterns of calibration variables contributing to the variance of weights vary across waves. As weights must include the same reference characteristics in all waves, some variables drive the variance strong in specific waves, variables were still included if they provided important information in other waves.

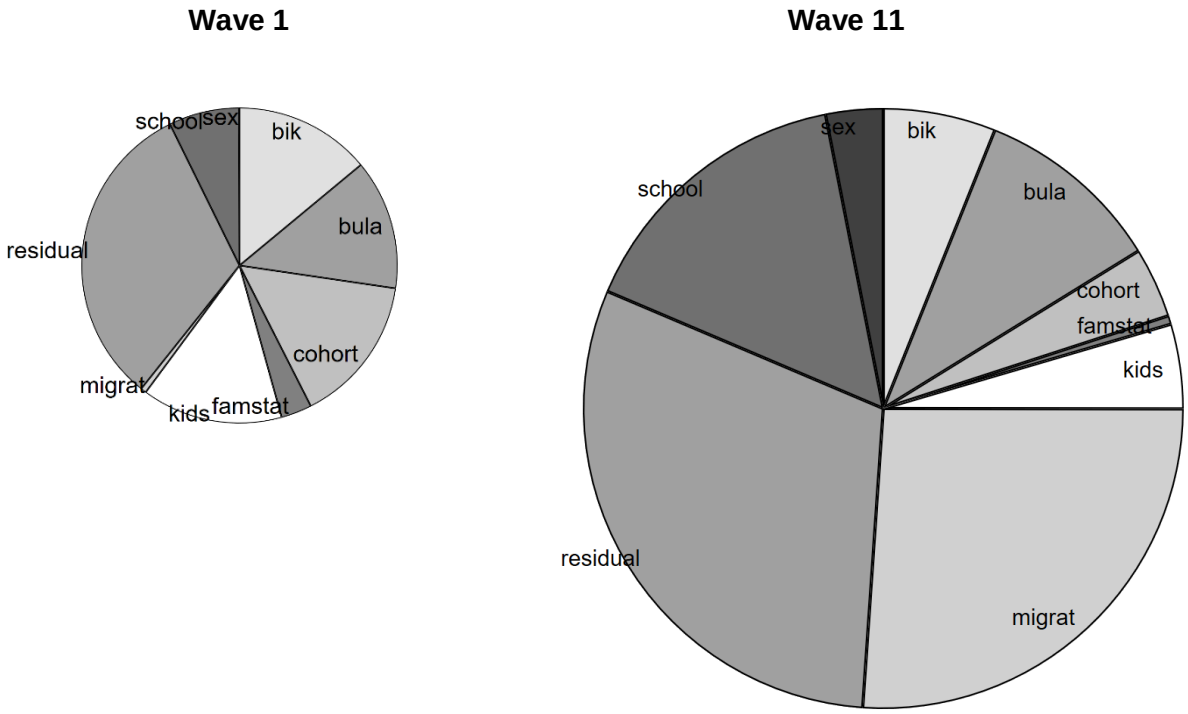

Figure 8: Analysis of variance of the cd2weight for wave 1 and 11

6 Comparing (un)weighted distributions with the Mikrozensus

This Section contains two example applications demonstrating how using weights can affect mean levels and deviations in the sample, and how comparable weighted results are with characteristics from the general population as measured by the Mikrozensus. The examples show the weighting effects for the number of children living in the household and labor force status. While the former has been considered as part of the calibration procedure, the latter is not included as a reference characteristic. First, a comparison of the unweighted with the Mikrozensus-based distribution exemplifies the patterns of bias in the unweighted data over the course of the panel. Next, a comparison of weighted (here: *cd2weight* for the entire sample) results with the characteristics from the population indicate the effect of weights.

Figure 9 displays the deviations of the unweighted and *cd2weight*ed data from the Mikrozensus for the mean level of the number of children living in the anchor respondent's household. Positive values indicate a higher proportion in the Mikrozensus (i.e., under-representation in

the pairfam data) and negative values over-representation in the pairfam data. The generated variable *nkidsliv* included in the anchor data sets (Brüderl, Garrett, et al., 2021) is grouped into three categories: “no children”, “one child”, and “2+ children”. The first graph shows the bias of the unweighted results for “no children”: Childless respondents of cohort 2 and cohort 3 are under-represented in the pairfam data, and this difference amounts to more than 10 percentage points for cohort 3 for most of the waves. While differences are less pronounced for respondents with one child, respondents (in particular of cohort 2 and 3) with two or more children are overrepresented in the pairfam sample. When using the weighting variable *cd2weight*, the distribution in the Mikrozensus is approximated nearly perfectly – a result which is not surprising, as *number of children in the household* is one of the variables used for the calibration process.

Further weighting examples using the number of children are included in the Quick Start do-file “Weighting” which is part of the pairfam Scientific Use File.

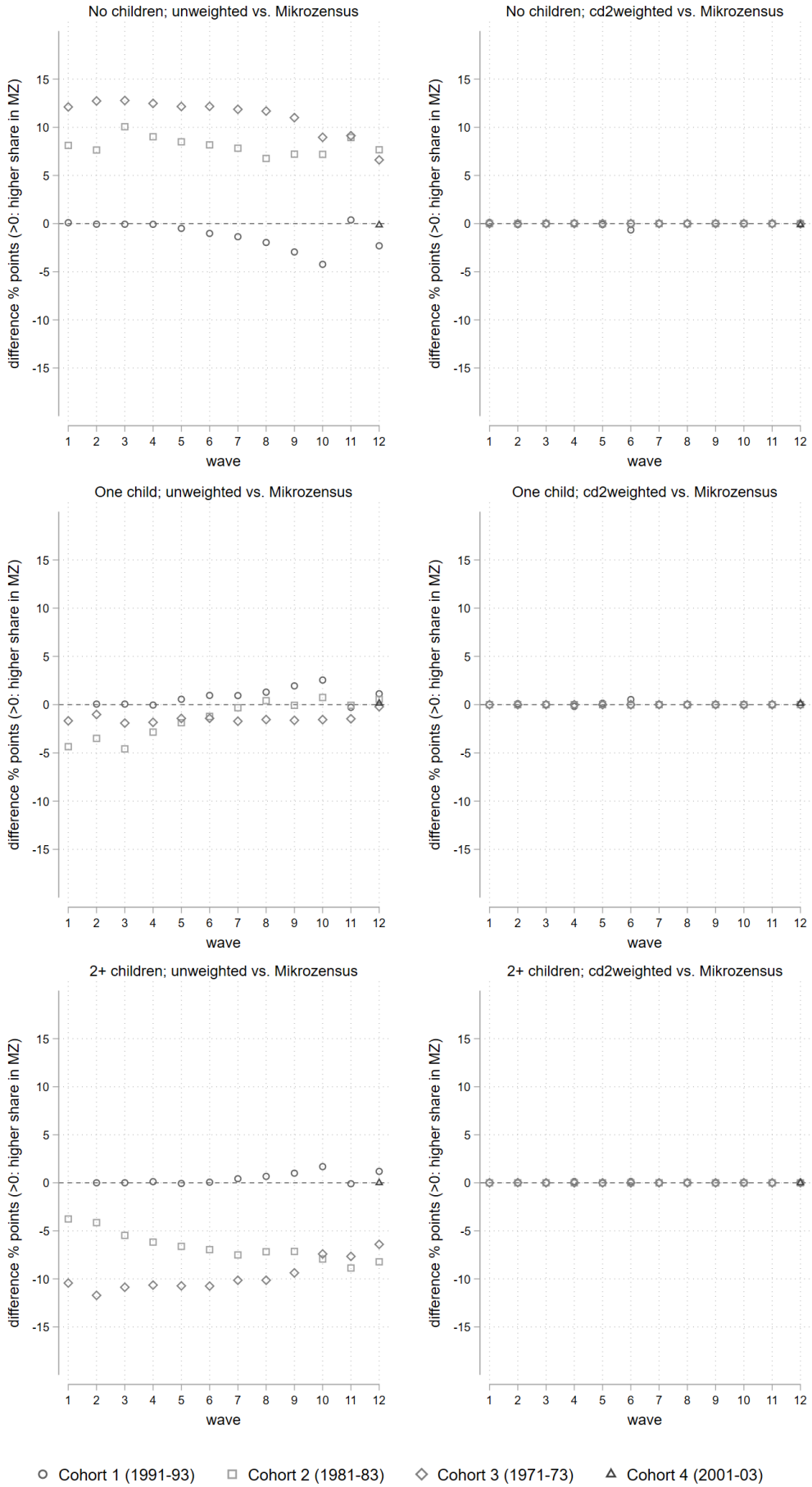

Figure 9: Difference between pairfam data (unweighted and cd2weighted) and Mikrozensus in number of children living in anchor respondent's household over panel waves

Do the calibrated design weights also help to approximate the distribution in the population for variables which have not been part of the weighting procedure? To examine this question, the anchor respondent's current labor force status is used in the second example. This is a variable that can be externally validated using information from the Mikrozensus, but has not been used to calculate the calibration weights.

In the pairfam anchor data sets, the generated variable *lfs* contains the anchor's labor force status (Brüderl, Garrett, et al., 2021). This 13-category variable was recoded here to the following six categories: "Full-time employment", "part-time employment", "unemployed", "other employment (e.g., marginal)", "in education", and "other non-employment", and these categories were then harmonized with the operationalization of the Mikrozensus (e.g., "vocational training" recoded to "full-time employment", "parental leave" recoded to "other employment"). The share of missing values is very low for this variable (< 1%), so no imputation procedure was implemented for this exemplary analysis – the missing values were simply excluded from the analysis.

Figure 10 displays the results for the three categories "full-time employment", "part-time employment", and "unemployed' with the y-axis representing differences to the Mikrozensus. For the part-time employed, the values are closer to zero in the weighted analyses (right-hand side graphs), while they do not differ much for the full-time employed and unemployed. These findings are interpreted as a reflection of both the strengths and limitations of the pairfam weighting approach: While the included reference characteristics do not help to tackle biases for full-time and unemployment, they do help to correct for proportions of part-time employment, as they are likely more strongly associated with the included reference characteristics such as gender, cohort, and region.

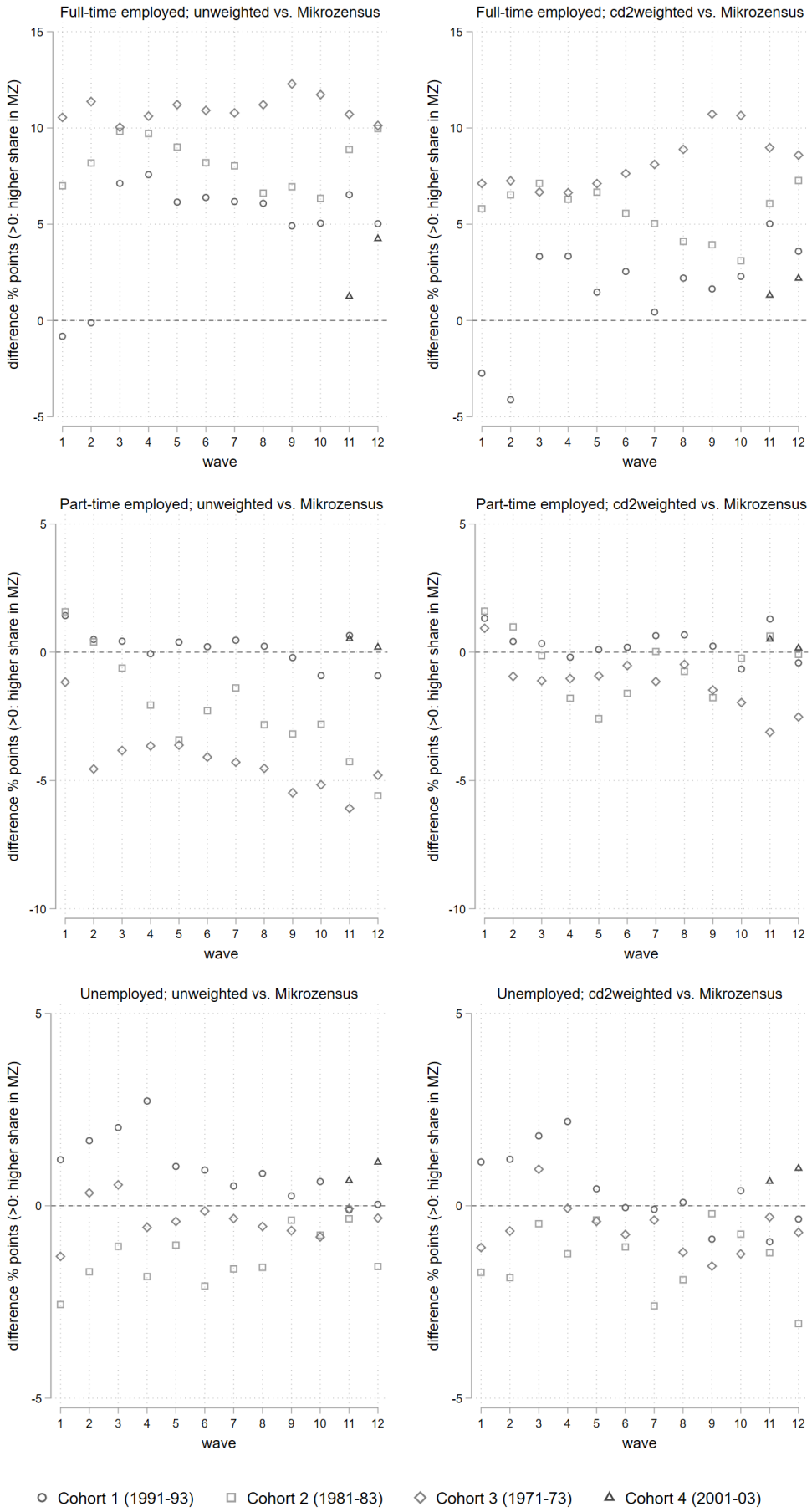

Figure 10: Difference between pairfam data (unweighted and cd2weighted) and Mikrozensus in anchor respondent's labor force status over panel waves

7 Recommendations

The survey design of the pairfam data is complex in terms of sample selection (i.e., by cohort, Eastern/Western Germany, and base/refreshment sample) and attrition patterns (Brüderl, Schmiedeberg, et al., 2021). Accordingly, the pairfam group recommends using the weights provided with the respective release, as the risk of misspecification by not addressing these issues outweighs the lower efficiency and lower statistical power often associated with weights (Bollen, Biemer, Karr, Tueller, & Berzofsky, 2016). However, a general discussion about whether or not to weight is beyond the scope of this report (see Bollen et al., 2016; Gelman, 2007; Solon, Haider, & Wooldridge, 2015).

For descriptive analyses, using weights is a suitable approach in most cases, as addressing selective participation is otherwise difficult to achieve. In regression models, if only one particular sample is of interest (e.g., pairfam base sample), controlling for cohort differences and for variables that predict selective drop-out might also be an appropriate approach. However, if analyzing more than one sample (pairfam base sample, DemoDiff, pairfam wave 11 refreshment) is intended, using calibrated design weights is highly recommended to control for various selection risks (i.e., by cohort, Eastern/Western Germany, and base/refreshment sample). All weights are part of each anchor data set so that they can be applied for each sample for each wave.

However, weights are not the cure for all selection issues. As socio-demographics (gender, migration background, educational attainment), regional differences (federal state, settlement structure), and family-related characteristics (relationship status, number of children) are explicitly considered in the computation of the calibration weights, they will provide accurate estimates for these variables. Accordingly, for research questions that require a selection-corrected sample for these variables (e.g., topics of separation, education-specific fertility, regional differences), weighted analysis are well-suited. Many other research questions might also benefit from the considered variables in the weights. For instance, although labor force status and income are not explicitly addressed in the weighting approach, selection processes tackled by the included reference characteristics likely cover major tendencies (see Section 6). Similarly, although norms and attitudes toward children and family are not included in the weights, weighted analyses account for selection by parenthood and relationship status and may therefore provide less biased results than unweighted analysis.

In contrast, the provided weights might be less suitable for some analyses, as weights will not reduce bias due to characteristics such as personality, health, intelligence, or religiosity unless they are related to characteristics corrected by the weights. For theses research questions, two approaches are suitable: Researchers may compute their own weights by predicting participation probabilities for each wave (starting with wave 2) using a selection of relevant

predictor variables. The provided calibrated design weight of wave 1 can serve as the first baseline weight.[6] Variables driving the selection process may also be included into the models. In this case, using the provided calibrated design weights is still recommended, as they control for sample design-based unequal inclusion probabilities and additional selection processes.

Please keep in mind that as the pairfam study observes distinct birth cohorts that may differ significantly in certain behaviors/characteristics, estimating point estimates (e.g., mean levels) over multiple cohorts might not be an adequate approach for most research questions. Often, presenting cohort-specific results might be a more informative choice. (Note that in contrast to calibrated design weights, design weights neither control for unequal cohort sizes in the net sample nor in the target population.) Weights are also appropriate if only specific subpopulations of the study are of interest (e.g., cohort 1991-93, residents of Eastern Germany, men, parents). Finally, even if weights are not being implemented in an analysis, the pairfam group recommends additional analyses with calibrated design weights as a robustness check. If the results are similar, this suggests that the original analysis does not suffer from (major) design, non-response, or attrition bias. If results differ, caution is advised.

[6] For more information on such an approach, read a former version of the pairfam Data Manual on weights (Release 11.0 or earlier). Such an approach may be adequate for research-specific weights, but is limited. For instance, calibration is done only at the first observation, which is adequate only upon the assumption that the population structure does not change over time.

Appendix

A.1 Distributions of calibrated design weights

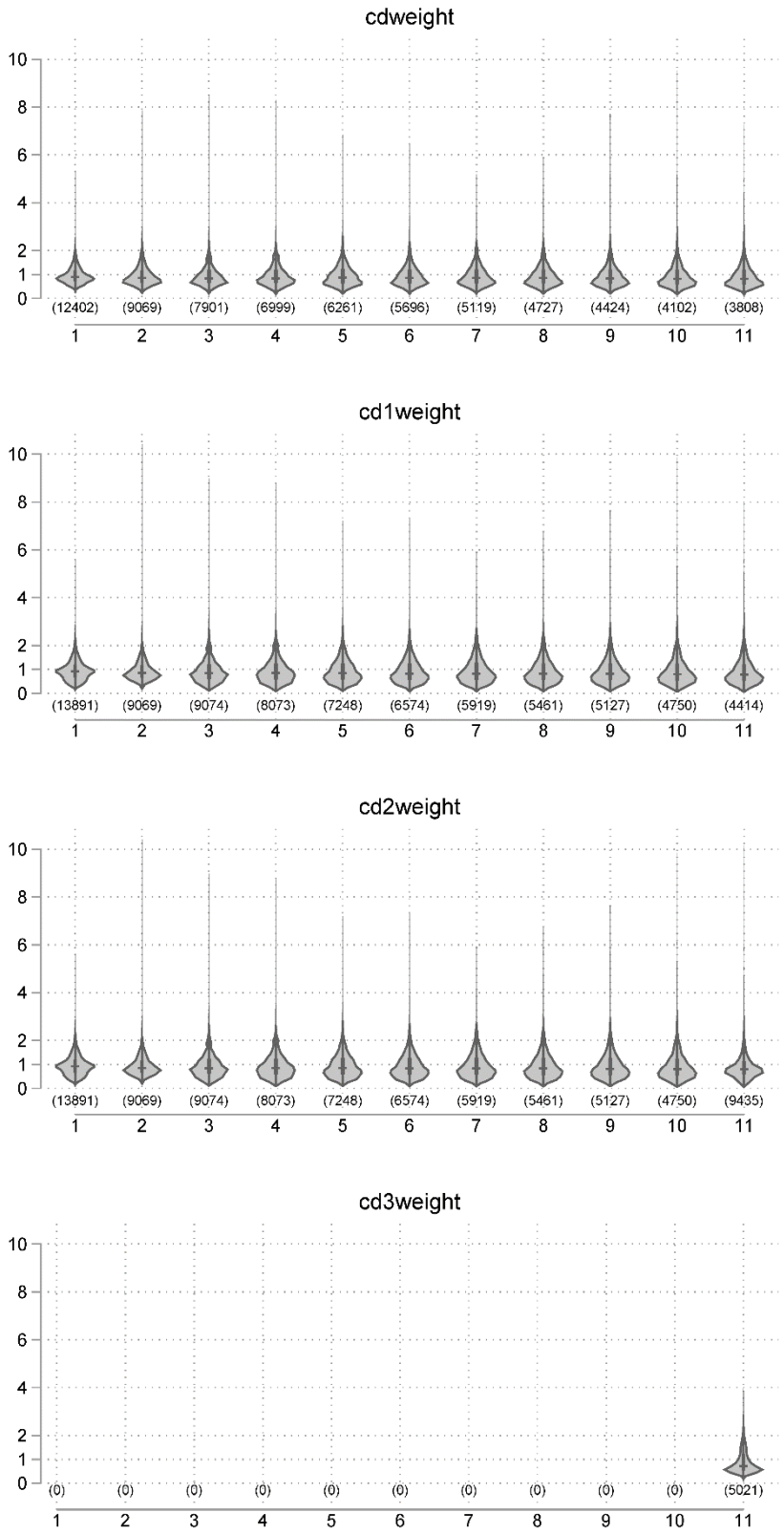

Figure 11: Violin plot showing the distribution of cdweight over waves

A.2 Means of calibrated design weights for reference characteristics over waves

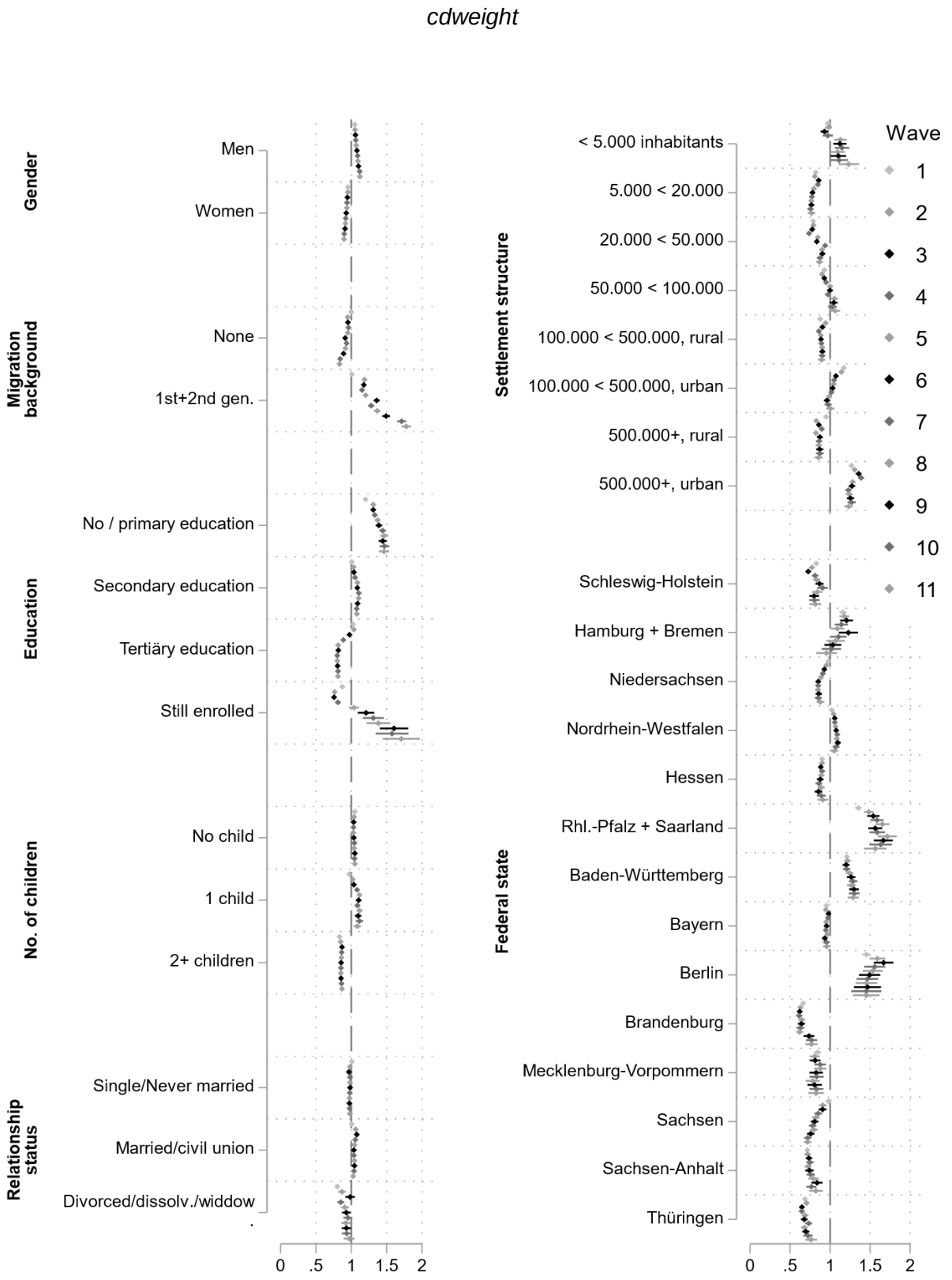

Figure 12: Means of cdweight for different subgroups over waves

cd2weight

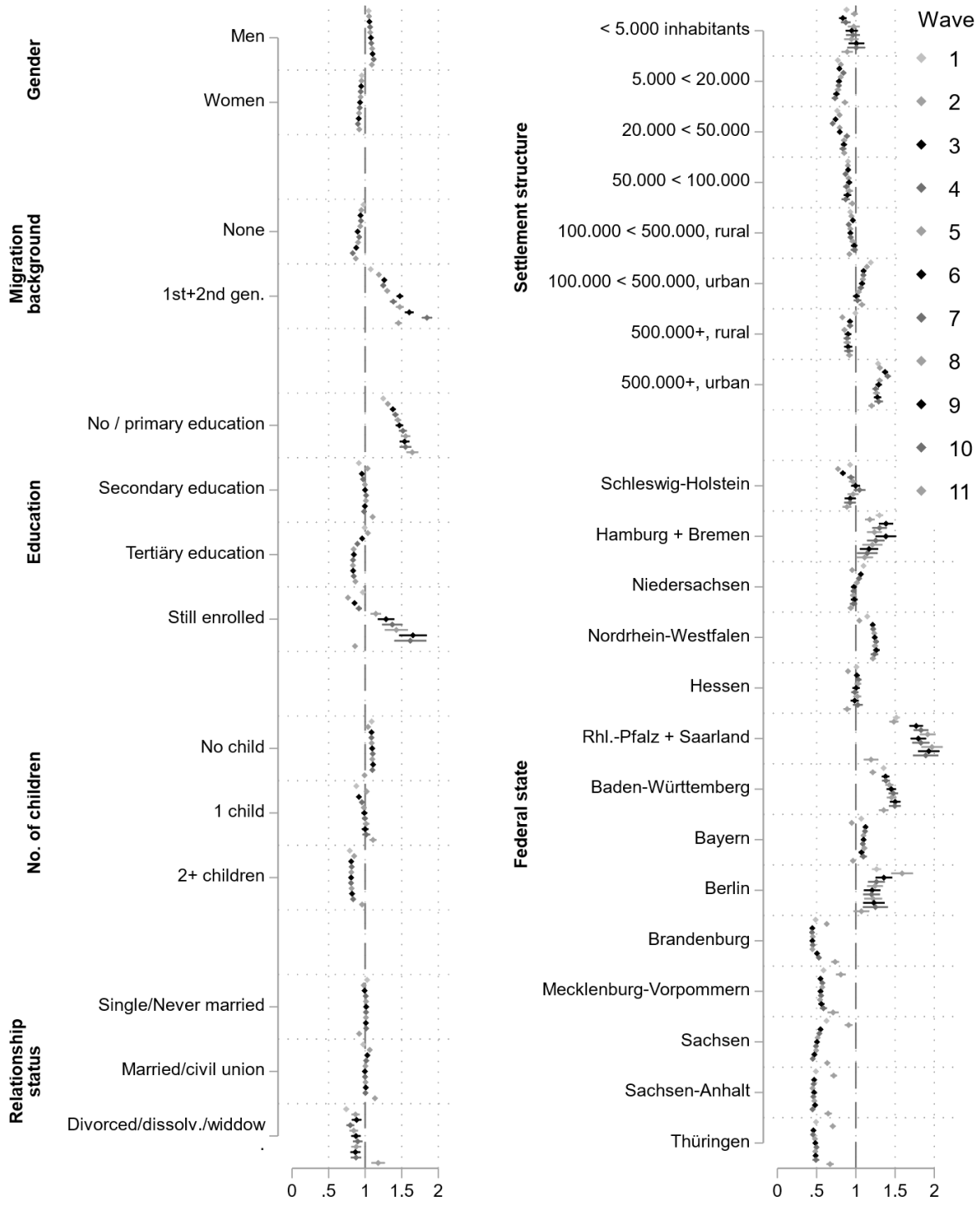

Figure 13: Means of cd2weight for different subgroups over waves

References

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Annual Review of Statistics and Its Application*, *3*(1), 375–392. https://doi.org/10.1146/annurev-statistics-011516-012958

Brick, J. M., Cervantes, I. F., Lee, S., & Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology*, *37*(1), 1–12.

Brüderl, J., Garrett, M., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., … Schumann, N. (2021). *pairfam Data Manual*. *LMU Munich: Technical Report* (Vol. 12.0). https://doi.org/https://doi.org/10.4232/pairfam.5678.12.0.0

Brüderl, J., Schmiedeberg, C., Castiglioni, L., Arránz Becker, O., Buhr, P., Fuß, D., … Schumann, N. (2021). *The German Family Panel: Study Design and Cumulated Field Report (Waves 1 to 12)* (pairfam Technical Paper No. 01).

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, *88*(423), 1013–1020.

Gabler, S. (2004). Gewichtungsprobleme in der Datenanalyse. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *44*(Sonderheft), 128–147.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, *22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Kolenikov, S. (2014). Calibrating Survey Data using Iterative Proportional Fitting (Raking). *The Stata Journal:*, *14*(1), 22–59. https://doi.org/10.1177/1536867x1401400104

Kolenikov, S. (2019). Updates to the ipfraking ecosystem. *The Stata Journal*, *19*(1), 143–184. https://doi.org/10.1177/1536867X19830912

Larsen, M. D., Qing, S., Zhou, B., & Foulkes, M. A. (2011). Calibration Estimation and Longitudinal Survey Weights : Application to the NSF Survey of Doctorate Recipients. *Joint Statistical Meetings - Section on Survey Research Methods*, 1360–1374.

Lavallée, P., & Beaumont, J.-F. (2015). Why We Should Put Some Weight on Weights. *Survey Insights: Methods from the Field*, *Weighting:* https://doi.org/10.13094/SMIF-2015-00001

Lohr, S. L., & Rao, J. N. K. (2000). Inference from Dual Frame Surveys. *Journal of the American Statistical Association*, *95*(449), 271–280. https://doi.org/10.1080/01621459.2000.10473920

Lundström, S., & Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, *15*(2), 305–327.

Müller, B., & Castiglioni, L. (2015). Stable relationships, stable participation? The effects of partnership dissolution and changes in relationship stability on attrition in a relationship and family panel. *Survey Research Methods*, *9*(3), 205–219. https://doi.org/10.18148/srm/2015.v9i3.6207

Müller, B., & Castiglioni, L. (2020). Do Temporary Dropouts Improve the Composition of Panel Data? An Analysis of “Gap Interviews” in the German Family Panel pairfam. *Sociological Methods and Research*, *49*(1), 193–215. https://doi.org/10.1177/0049124117729710

Sand, M. (2018). *Gewichtungsverfahren in Dual-Frame-Telefonerhebungen bei Device-Specific Nonresponse* (GESIS-Schriftenreihe No. 20). GESIS - Leibnitz-Institut für Sozialwissenschaften, Köln. https://doi.org/10.21241/ssoar.60293

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What Are We Weighting For? *Journal of Human Resources2*, *50*(March), 301–316.

Suckow, J., & Schneekloth, U. (2009). Beziehungen und Familienleben in Deutschland (2008/2009) Welle 1. *Methods Report. TNS Infratest*, *München*.