methods, data, analyses | Vol. 18(1), 2024, pp. 79-108 DOI: 10.12758/mda.2023.06

Challenges in Assigning Panel Data With Cryptographic Self-generated Codes – Between Anonymity, Data Protection and Loss of Empirical Information

Christina Beckord

ehs University of Applied Sciences for Social Work, Education and Nursing, Dresden

Abstract

The assignment of questionnaires between the 13 survey waves in the panel study "Crime in the Modern City" (CrimoC) was done by matching self-generated codes. This method was challenging because the individual codes tend to be ambiguous, prone to errors and the resulting panel data can be biased. The individual data were merged over time using an error-tolerant matching process with manual handwriting comparison. Despite these problems, there is no alternative to the chosen method with regard to anonymity and data-protection. Until now, the self-generated codes of each new survey wave were matched against the codes of the last and second-last wave. Over the years, this led to an increasing discrepancy between the data originally collected and the data linked to the panel. For this reason, first in a pretest and later for the complete sample, the cases that had not yet been linked to the panel were subsequently matched with earlier waves. This panel consolidation proved to be very successful. A total of 3,589 original missing units were subsequently filled with case data. This paper describes the steps taken to optimize the quality of the panel data set and illustrates exemplarily on specified criteria which properties of the panel data set could be improved. Since the importance of panel studies is steadily increasing in social science research this paper is relevant for researchers who need to make matching decisions within panel studies. Assurance of anonymity can counteract panel attrition. Self-generated codes represent one possibility in this regard, and are discussed in terms of feasibility and effectiveness.

Keywords: panel data, missing unit, personal codes, assignment rates

Longitudinal and panel designs are useful for analyzing intra- and inter-individual changes. A major challenge in this context is the matching of individual data over time. If no data from a new survey time point can be assigned to a previous time point, this can have two causes: Either the person did not actually participate in the new survey (refusal) or he or she did participate but the data could not be linked to previous data. Both cases lead to missing data in the panel data set: the so-called wave nonresponse or missing units.

Probably the simplest and safest matching method is to use participants' plain names. This, however, has the decisive disadvantage that the participants cannot be assured of the anonymity of their information, which can lead to refusals to participate, especially when sensitive content is being surveyed, as in the example of juvenile delinquency used here. In addition, the initial population of the reported study "Crime in the Modern City" (CrimoC) consisted of pupils aged 13 on average who attended a school in the city of Duisburg in 2002 (see Bentrup, 2019). Thus, a data protection concept also had to be developed due to the young age and the associated necessary declaration of consent by the parents. Together with the State Commissioner for Data Protection and Freedom of Information of North Rhine-Westphalia, it was decided to use self-generated personal codes that would allow the individual data to be combined while guaranteeing anonymity. This procedure was chosen for two interrelated reasons: first, to grant respondents the anonymity of their answers, and second, to make any possibility of de-anonymization by third parties impossible, since, violations relevant to criminal law were inquired about. These individual codes are self-generated by each respondent through responses to 6 to 10 targeted questions on time-stable characteristics (Pöge, 2008: 60). To ensure good reproducibility, letters from own name or the name of close relatives are often used. The goal is to obtain combinations that are as unique as possible. Over a total of 13 survey waves, this procedure proved to be a stable allocation procedure for most participants. Nevertheless, at each point in time, it was not possible to link a certain proportion of participations to the panel dataset. For this reason, the missing units were composed of individuals who either did not participate or did participate, but the individual data could not be matched to the panel data set using the self-generated code. It is precisely in this last case that the described data optimization comes into play. The panel consolidation describes a procedure with which missing units are subsequently replaced by originally collected data. The question that arises after such a time-consuming and challenging process whether the new data situation represents an improvement over the original panel.

Direct correspondence to
Christina Beckord, ehs University of Applied Sciences for Social Work, Education and Nursing, Dresden
E-mail: christina.beckord@ehs-dresden.de

While there are possibilities to address missing values at the statistical level (Rubin, 1987; Reinecke & Weins, 2013; Kleinke, Reinecke & Weins, 2021), even these methods have their limits. For this reason, the stated goal should be to integrate as many cases into a panel dataset as possible. For example, it is not possible to impute outstanding events that are not influenced by any predictors. One such example are typical stages of life such as starting an own family. For a sufficient subgroup analysis with longitudinal data as much cases of the data collection as possible should be included in the panel data. In addition to certain subgroups, an existing bias in the linkage by certain characteristics, e.g., gender, also poses a difficulty in interpreting the results. But does the consolidation call into question the quality of the previous panel data set? For this purpose, the main variable "juvenile delinquency" is examined in more detail below. If there are no changes in this characteristic in longitudinal analysis, this would indicate that the new cases compared to the already matched cases are at random regarding the dependent variable.

All in all, the described panel consolidation is considered a success if dropout from relevant subgroups can be minimized, biases in the panel data set can be reduced, and at the same time the structure of main dependent variables (here: juvenile delinquency) do not change from the original data set.-

Therefore, this paper begins with a description of (1) the starting point – the original CrimoC-data, the application and limitations of the self-generated codes and (2) the performed optimization of matching cases within the existing 13-wave panel data. Furthermore, it is (3) defined when the panel consolidation is considered successful and (4) what improvement could be achieved by the newly connected cases.

The Starting Point: "Crime in the Modern City" (CrimoC)

Crime in the Modern City (CrimoC) is a prospective panel study that began in 2002, surveying 7th grade pupils from public schools in the German city of Duisburg. The self-report questionnaire had the goal of explaining and monitoring the emergence and development of deviant and delinquent styles of behavior throughout the phase of adolescence (Sedding & Reinecke, 2017; Reinecke et al., 2015). As possible causes of these phenomena, the study focuses not only on structural conditions and processes on the macro-level but also on the meso- and micro-level (e.g., social milieus, moral orientation, lifestyle, how spare time is spent, attitudes, norm orientations, social environment; detailed information about the study can be obtained from the webpage www.crimoc.org; Boers et al., 2010; Boers & Reinecke, 2019). Due to the satisfying re-interview rates, and the successful panel construc-

tion, the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) has extended its funding in three-year intervals up to now.

Data Collection

The longitudinal self-report panel design evoked three major challenges: (1) respondents' retrieval after age-related school leaving despite the assurance of anonymous answers, (2) the necessity of different data collection modes, and (3) the matching of individual data over time by simultaneous assurance of response anonymity.

At the beginning of the CrimoC-study, the researchers attempted to collect data from all 7th graders in all public schools of Duisburg, an industrial city in the Rhine-Ruhr-Area with a long tradition in coal mining. In Germany, there are five different types of schools that follow elementary school: the *Hauptschule*, a school with a lower level of education which ends after grade 9, the *Realschule*, a medium-level school which ends after grade 10, the *Gymnasium*, the highest educational level which ended for our cohort after grade 131, the *Gesamtschule*, a combination of Realschule and Gymnasium which enables more pupils to achieve a higher educational level, and the *Förderschule* where pupils with learning disabilities receive special support. Of all 56 schools of Duisburg, 16 refused to participate. The progress of data collection was adjusted to the age and life stage of the respondents. From age 13 to 20, the survey was conducted annually, and from age 20 to 30, every two years (figure 1, in detail, see Bentrup, 2019).

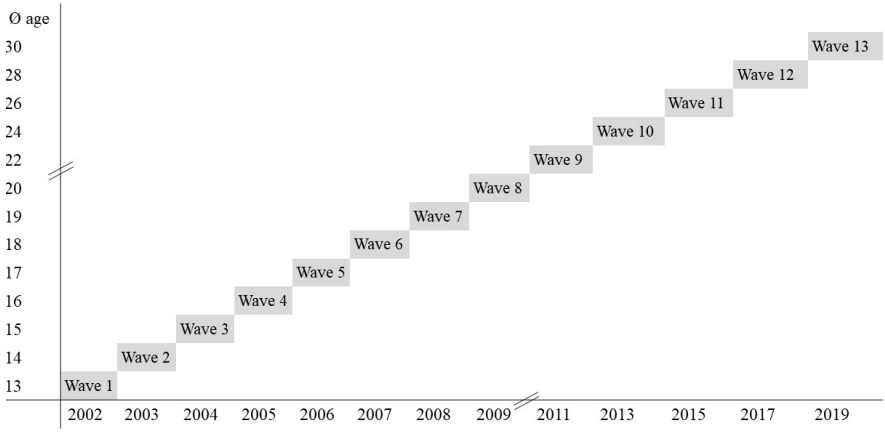

Figure 1 CrimoC survey design

1 Meanwhile the Gymnasium and Gesamtschule end after grade 12.

The first four waves took place in the school context with self-administered paper-pencil questionnaires, while the following waves of data collection were used for a stepwise change into postal mode. In order to contact the participants after leaving school, all respondents were asked to share their addresses (independent of the completed questionnaire). The resulting contact database was updated within each of the following data collections. If participants changed their residence, they had the possibility to communicate the new address to the project team via the project's webpage. Respondents who did not report their new residence were searched in local registers of residence. After every postal data collection, an additional personal contact phase was carried out for all contacts in the database who did not participate in the actual wave. This could be executed despite the assurance of anonymity because all participants filled out a separate address card to receive an incentive of 25 Euros for their participation. In this case, interviewers contacted respondents to motivate them to participate after all. If the respondents agreed, they were given a questionnaire by the interviewers (if necessary), which was to be completed without the presence of the interviewers and later collected again by the interviewers. In this way, 13 waves could be realized in 18 years.

The complex study design necessitates a closer look at the different datasets. First, one has to distinguish between different terms: the cross-sectional datasets for each time point t (CS_t), which include all individuals who filled out a questionnaire during a data collection wave. Second, there are the matched individual data – the 13-wave panel dataset that includes all cases with at least one match to another time point. The single cases it this panel dataset differ regarding the number of participation (independent of whether this missing unit is due to non-participation or not being matched to the dataset). The possible data range is between 2 and 13 points in time or in other words, the number of missing units varies from 0 to 11. The largest number of cases per time point is therefore obtained when all missing units are tolerated. This number of cases in the panel per time point (t) is referred to below as panel-cross-sectional data[2] (PCS_t). Additionally, there are the complete panel datasets, which contain only those respondents who have participated any time during the period of interest, and which could be successfully matched to the previous individual data ($P_{t1\text{-}t13}$). Fourth, one can use panel data sets with missing units, which include all cases with a tolerated number of missing units ($P_{txi,\ txj,\ldots,tX}$).

2 Even though strictly speaking it is the number of 2-wave panels from t to t+1 or t-1.

The study started in 2002 with a survey of the initial population of 3,411 7th-grade pupils in Duisburg. In the following years, the cross-sectional re-interviewing rates ranged between 85 and 92%[3].-

Previous Matching of Individual Data Over Time

In order to enable the questionnaires from the different survey waves to be assigned, individual codes were used which were requested via code sheets. In the course of the interview, each respondent filled out a code sheet containing five or - from the 2003 survey onward - six personal questions, the answer to each of which represented a specific letter or number and was to be noted down accordingly. The questions referred to unchangeable characteristics of the respondent or his environment (natural hair color, name of father, etc.). This letter-number combination finally formed the entire code. In each survey wave, the code was filled in by the participants at the beginning of the survey. Since the codes in each survey contained the same information, the codes filled out by the same person in the different waves should have to be identical.

The questions to create the code included:

Co001: The first letter of the father's first name

Co002: The first letter of the mother's first name

Co003: The first letter of your first name

Co004: The two-day digits of your own birthday

Co005: The last letter of the own hair color

Co006: The last letter of your own eye color

Since 2009 additional:

Co011: The last letter of own surname (in case of name change, the birth name)

Since the survey year 2003, the following questions have also been asked:

Co007: Survey participation in the previous year (yes/no)

Co008: Change of school in the past year (yes/no)

Co009: Not transferred in the past year (yes/no)

---

3 In detail: 2003 n = 3,392; 2004 n = 3,339; 2005 n = 3,243; 2006 n = 4,548; 2007 n = 3,336; 2008 n = 3,086; 2009 n = 3,090; 2011 n = 3,050; 2013 n = 2,850; 2015 n = 2,754, 2017 n= 2,778; 2019 n= 2,697. The data collection in the year 2006 was the most challenging one. Due to the school leave of respondents in the lower educational level schools and the compulsory school attendance for all adolescents up to age 18, the attempt was made to retrieve these school leavers in selected classes at vocational schools. A consequence was that the cross-sectional data includes additional cases of individuals who attended these classes but who did not participate before. These additional cases leave no impact on the panel-dataset because they could not be matched to previous cases.

Co004, Co007, Co008 and Co009 have been included in the code sheet since the year 2003. In addition, the design of the code sheet has been changed. In 2002, respondents had to provide their respective answers to the code questions in a box in handwriting; since 2003, all possible letters have been shown as answer options to be checked off (see appendix A).

Since the fifth wave of the survey (2006), Co008 has not been collected due to the end of the school career of most respondents. Co009 has been collected since 2006 only for those respondents who attended a Gymnasium or a Gesamtschule. Since the eighth wave of the survey (2009), only survey participation in the previous year (Co009) was asked. In addition to the six code questions and the supplementary questions, information on the respondent's gender and the (most recent) school attended was available for the questionnaire assignments.

The function of the code requires that the codes are a) unique, i.e., that the individual parameters have enough variance so that the codes can be uniquely identified (identification), b) that the participants answer the individual code questions exactly the same over time (replication), and c) that the queried characteristics are indeed time-invariant characteristics.

In 2002, the problem of identifying individual data over time was posed by multiple occurrences of the same complete codes. By adding one code question (the last letter of one's first name), the uniqueness of the code could be greatly increased. In 2002, there were 324 double occurrences of the complete code (7.9%), 18 triple occurrences (0.5%), and 5 quintuple occurrences (0.1%); in 2003, the six-digit code had only 32 double occurrences (0.9%) (cf. Pöge, 2007: 6; Pöge, 2008: 62). This figure remained between 2.0% in 2006 and 0.1% in 2009 across all subsequent survey waves. In principle, respondents were willing to fill in the code with an overwhelming majority (98.5% in 2005 to 99.6% in 2006).

However, the problem of replicating the individual codes remained. For this reason, an error-tolerant matching procedure was developed in which gradually more and more errors in the code were allowed (cf. Pöge, 2005: 66). To provide additional certainty about the matching, each potentially matching questionnaire from two points in time was subjected to a manual handwriting check.

The steps of the error-tolerant matching procedure are hierachical and allow more variation in the code with each step. Accordingly, the assignment rate decreases with each additional step (table 2). Each step consisted of two sub-steps to keep the number of reconciliations to be performed manageable: first, gender and school attended had to be compared in addition to the codes (with the number of errors allowed in each case). Moreover, students were matched on the basis of additional variables (Co007, Co008, Co009), which asked whether they had participated in the survey in the last year, as well as whether they had changed schools or stayed behind. Second, the additional conditions to be fulfilled were successively relaxed and in some cases omitted altogether. In this way, there were controls for

0-3 errors in the code and the release of the control variables, so the basic structure was a 4*2 pattern.

After each step, the matched questionnaires were then subjected to a manual handwriting check. This check was performed for several reasons: on the one hand, there was the possibility that individual codes were not unique (especially when tolerating errors) On the other hand, it was an additional control on the basis of the handwriting style and/ or similarities in the content in the questionnaire. Those pairs of questionnaires that had obviously been completed by the same person were removed from the data sets so that they were no longer available for the subsequent matching steps. Non-matching questionnaires remained in the data sets, possibly to be identified as matching in one of the next matching steps.

The 2007 and 2008 data collections will serve as an example of the chosen approach (table 1). The respective cross-sectional data comprised n=3.336 in 2007; n=3.086 in 2008 (Daniel & Erdmann, 2017: 8).[4]

It can be seen that the number of comparisons increases as the error tolerance increases, whereas the assignment rate decreases. A total of 4,407 potentially matching pairs of questionnaires were identified, of which 2,698 (61.2%) were found to be matches during the handwriting checks. In terms of the cross-sectional data set of 2008, this means that of the 3,086 cases available, 2,698 (87.4%) could be linked to the cross-sectional data of the previous wave.

In addition, controls were also conducted between survey waves that were not directly consecutive (figure 2). The first four survey waves were fully matched. For economic reasons, starting with the fifth survey wave in 2006, the codes of a cross-sectional data set, which had not yet been assigned to the panel after the matching with the immediately preceding wave described above, were compared with the unassigned codes from the penultimate wave.

Between these data, in a first step in which the condition of fully matching codes and fully matching additional variables were checked, 1,403 potentially matching pairs of questionnaires were identified. 1,343 (95.7%) of these were found to be matches in the subsequent handwriting checks. These were marked as matches and removed from the two cross-sections for further matching. The control steps shown in table 2 followed in order.

Since the matching was based on the cross-sectional data, these cases were matched to the existing panel data set in a next step. This again reduced the number of cases, so that in the previous example, the original PCS (oPCS) for 2008 included a total of 2,412 cases (Erdmann, 2021).

4 The original table was summarized to the 4*2 steps described above for illustrative purposes, even though a total of 10 steps were performed in the matching process to keep the size of each list to be compared manageable.

Table 1 Performed control steps 2007/2008

Step	Codevariables	Additional variables
S1	without errors	without errors
S2	without errors	no restriction
S3	one error	without errors
S4	one error	no restriction
S5	two errors	without errors
S6	two errors	only selected restricitons
S7	three errors	without errors
S8	three errors	only selected restrections

Table 2 Number of checks and matches

		Number of checks	Match		No match	
Errors	Step	n	n	%	n	%
Without errors	1	1,403	1,343	95.7	60	4.3
	2	584	506	86.6	78	13.4
One error	3	415	370	89.2	45	10.8
	4	371	258	69.5	113	30.5
Two errors	5	138	104	75.4	32	24.6
	6	1,190	89	7.5	1,101	92.5
Three errors	7	194	24	12.4	170	87.6
	8	112	4	3.6	108	96.4
Total	1-8	4,407	2,698	61.2	1,709	38.8

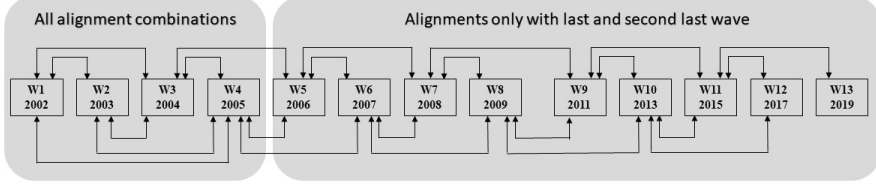

Figure 2 Matches performed as part of the original panel construction

Discrepancies Between Cross-sectional and Panel Data

Consequently, as was made clear in the previous section, there is a discrepancy in the number of cases between CS_t and $oPCS_t$. Table 3 shows the differences in the number of cases between the cross-sections and the associated panel cross-sections, as well as the differences between CS_t and $oPCS_t$. Two things become clear: The first four waves, which were fully matched, show the best assignment rate. The increased difference between CS_t and $oPCS_t$ in the first survey are due to the shorter code, the lack of additional questions, and the more difficult layout of the code query (see previous section). All other data collections show a much larger difference between CS_t and $oPCS_t$. Ideally, this drop out is at random, i.e., does not exhibit systematic failures.

In summary, it can be stated at this stage that since the sixth survey wave in 2006, between 21.8% and 38.1% of the participating individuals could not be assigned to the panel. However, since contact data are available from all individuals to ensure the postal survey, it should theoretically be possible to assign them to the panel data set.

In addition to the loss of cases, the linkage to the panel exhibits additional biases. In the earlier waves (w1-w4), these relate to the type of school. High school students are more strongly represented in the panel than in the cross-section. For all waves, there is a clear bias with respect to the gender of the respondents; female participants are significantly overrepresented in the panel data set (counts are in table 8).

Finally, due to the general loss of cases, some subgroups of special interest were significantly reduced. This reduction becomes more pronounced the later the point in time considered for the identification of a subgroup in the dataset (e.g. parenthood). For instance, in 2011, 168 of the respondents reported having at least one child of their own. Of these, however, only 106 were found in the oPCS of the year 2011. For this reason, a pretest in 2011 attempted to link additional parents to the panel by performing code matches for survey periods more than two time points apart. The developed method turned out to be surprisingly successful. Further matching increased the number of parents in the consolidated PCS_t ($cPCS_t$) by 48 cases to a total of 154, representing 92% of the parents in the CS. Due to its success, it was decided to apply this procedure to all cases of the full panel. The procedure is explained in the next section.

Table 3 Case numbers of data sets in cross sections, panel cross sections and their difference

	2002	2003	2004	2005	2006	2007	2008	2009	2011	2013	2015	2017	2019
CS_t	3,411	3,392	3,339	3,243	4,548	3,336	3,086	3,090	3,050	2,849	2,754	2,778	2,697
$oPCS_t$	2,750	3,132	3,177	3,195	3,032	2,587	2,412	2,304	2,100	1,912	1,812	1,760	1,670
$oDiff_t$	661	260	162	48	1,516	1,026	674	786	950	937	942	1,018	1,027
%	19.4	7.7	4.9	1.5	33.3	30.8	21.8	25.4	29.7	32.9	34.2	36.6	38.1

CS_t: cross-sectional data for time point t; $oPCS_t$: original panel data for time point t; $oDiff_t$: CS_t-$oPCS_t$; %: $oDiff_t/(CS_t/100)$

Improvement in Assignments – The Panel Consolidation

First of all, it should be summarized to which criteria the improvement of data quality should be determined. Four criteria are applied in this paper:

1. *Increase in the number of cases per time point in the panel*: the initial aim is to replace as much missing units as possible through additional matching with aubsequent allocations. This means that the consolidated PCS_t ($cPCS_t$) have a larger number of cases than the corresponding $oPCS_t$.
2. *Reduction of socio demographic bias:* since all cases that have not yet been allocated originate from the CrimoC-population, the overrepresentation of females should decrease within the consolidation, since more data from male respondents should be matched.
3. *Improving the number of cases of relevant subgroups:* As parents are an important subgroup for a follow-up project, the difference in the number of cases between cross-sectional and panel data should be reduced.
4. *No changes in the structure of the dependent variable:* the longitudinal structure of the main dependent variable (juvenile delinquency) should not change significantly. Otherwise this would be an indicator for a relevant bias in the previous panel construction and related to this in the interpretation of previous results.

The Consolidation Procedure

The procedure was analogous to the original panel construction. First, the cases of the cross sections were selected that could so far not be matched to the panel ($oDiff_t$ in table 3). For each of the 29 potential additional checks listed in table 4, the cases of the oPCSt were selected that so far have a missing unit for the wave of interest. For example, the 2007 cross-section was reduced to those cases that were not previously part of the 2007 panel cross-section. For the matching with the 2004 panel wave, the 2004 panel cross-section was reduced by the cases that already had a link to 2007. For the resulting two partial data sets, SQL queries were run in *Access* to identify identical codes or, in the context of the error-tolerant procedure, the corresponding potential matches.

Because these subsamples were considerably smaller than had been the case in previous panel checks, matching was performed in two steps: Step 1 included all cases with identical codes for each match, and the additional variables were not equated. This corresponds to S2 of the original panel controls (table 1).

Table 4 Potential for panel consolidation

Survey year	Already performed checks	Potential further checks
2006	2005, 2004	2003
2007	2006, 2005	2004, 2003
2008	2007, 2006	2005,2004, 2003
2009	2008, 2007	2006, 2005, 2004, 2003
2011	2009, 2008	2007, 2006, 2005[1)]
2013	2011, 2009	2008, 2007, 2006 2005
2015	2013, 2011	2009, 2008, 2007, 2006
2017	2015, 2013	2011, 2009, 2008, 2007
2019	2017, 2015	2013, 2011, 2009, 2008

1) In the first comparisons, it turned out that the complete comparison of waves 1 to 5 already performed meant that further checks in these waves for later points in time were not very successful. For this reason, an additional comparison with 2004 was not performed in 2011.

In the second step, one error was tolerated in the code, and the additional variables remained unrestricted (S4 of the original panel controls). Further checks were deliberately omitted because manual handwriting comparison, which became of increasing importance especially for assignments with more than one tolerated error, becomes increasingly difficult over a greater temporal distance.

The number of reconciliations is summarized for each survey wave in Table 5; a detailed list of all reconciliations per survey wave can be found in appendix B. A total of 7,068 potential matches were checked, of which 3,589 (50.78%) resulted in new matches in the existing panel dataset. It is important to note here that the aim was not to link new cases to the dataset, but to fill gaps (in the form of missing units) through subsequent checks, i.e., the total number of cases before and after panel consolidation is identical at 4,076 cases (last row table 5). The table also illustrates that the number of matches, as well as the assignments found, increased with distance from the starting point of the study, the fully controlled five-wave panel. Appendix C illustrates two typical cases of the consolidated complete panel data set. A detailed documentation of the occurring errors by code question does not exist, as the queries since 2003 have been carried out and documented by number of errors, but not broken down by code question.

Table 5 Panel consolidation checks and matches

Aligned wave	Number of checks	New matches	Matches (%)	$oPCS_t$ (n)	$cPCS_t$ (n)	Increase (%)
t_5 2006	169	5	2.96	3,032	3,037	0.16
t_6 2007	428	123	28.74	2,587	2,710	4.75
t_7 2008	665	333	50.08	2,412	2,745	13.81
t_8 2009	815	459	56.32	2,304	2,763	19.92
t_9 2011	775	468	60.39	1,812	2,324	28.26
t_{10} 2013	976	480	49.12	1,912	2,392	25.10
t_{11} 2015	1,115	512	45.92	1,812	2,324	28.26
t_{12} 2017	1,058	597	56.43	1,760	2,357	33.92
t_{13} 2019	1,067	612	57.36	1,670	2,282	36.65
total	7,068	3,589	50.78	4,076	4,076	100.00

$oPCS_t$= original panel cross-sectional data set; $cPCS_t$= consolidated panel cross-sectional data set; Matches (%): new matches/ (number of checks/100); Increase (%) = $cPCS_t$/ ($oPCS_t$/100).

The 3,589 new matches are distributed among 1,071 participants, for whom one missing unit could be filled in 259 cases, two in 195 cases, three in 149 cases, four in 169 cases, five in 102 cases, six in 98 cases, seven in 73 cases, and eight original missing units could be replaced in 26 cases.

The greatest improvement was achieved for panel data sets with four to six missing units. Here, panel consolidation increased the number of cases by more than 500. But also the panel data sets with fewer missing units could be increased considerably. The 79 closed gaps for the continuous panel (first row table 6) are astonishing because, actually, comparisons were always carried out between three consecutive survey dates. Thus, a complete control was available for these cases. This may be due to three reasons: 1) in the handwriting control, a case was originally declared as non-matching but now declared as a match; 2) in the handwriting control, a questionnaire could not be found; 3) more than one gap was closed for some cases, so that there may be an increase in the number of cases for the continuous panel. The first possibility applies to 14 of the 79 new cases in the continuous panel dataset, and the second reason is crucial for 65 of the 79 cases: two missing units were filled with data for five of the cases, three missing units for one case, four missing units for nine cases, five missing units for 15 cases, seven missing units for 13 cases, and eight missing units for 16 cases. These cases were randomly tested for plausibility of assignment.

Table 6 Number of cases of original and consolidated panel data set by missing units

	oPCS		cPCS		Increase
Missing units	n	%	n	%	n
0	735	18.0	814	20.0	79
0-1	1,230	30.2	1,404	34.4	174
0-2	1,542	37.8	1,834	45.0	292
0-3	1,749	42.9	2,161	53.0	412
0-4	1,965	48.2	2,466	60.5	501
0-5	2,143	52.6	2,647	64.9	504
0-6	2,316	56.8	2,835	69.6	519
0-7	2,497	61.3	2,983	73.2	486
0-8	2,815	69.0	3,145	77.2	330
0-9	3,163	77.6	3,376	82.8	213
0-10	3,550	87.1	3,629	89.0	79
0-11	4,062	99.7	4,063	99.7	1
0-12*	4,076	100.0	4,076	100.0	0

% oPCS= n_{oPSC}/(4,076/100); % cPCS= n_{cPCS}/(4,076/100).
* 12 missing units are 14 (oPC_t) and 13 (cPC_t) cases, respectively, which were assigned to another time point, but the second case was classified as not qualitatively usable.

With regard to the first criterion for the improvement of data quality in the panel dataset - i*ncrease in the number of cases per time point in the panel* - it can be summarized that the number of cases in the cPCSt increased significantly compared to the oPCSt at all points in time. The later the time of the survey and thus the more additional comparisons were possible, the more missing units could be filled with empirical information.

Improvements in Content Due to the New Assignments

Following the encouraging results of the panel consolidation, the question arises as to its significance for the data structure. Based on the cross-sectional data, the quality of the assignments before and after panel consolidation can be assessed in terms of content to examine the quality criteria 2 to 4.

Examination of the Quality Criteria at the Content Level

Reduction of socio demographic bias: Table 7 illustrates the gender differences between the CS and oPCS. Within the oPCS, all time points are characterized by a higher proportion of female participants. If the panel consolidation meets the quality criterion, the difference between the proportion of females between the consolidated panel and the cross-sectional data should be smaller than between the original panel dataset and the cross-sectional data (cDiff % < oDiff %). Although the proportion is still higher than in the cross-sectional data all points in time of the consolidated panel meet this criterion.

Table 7 Gender differences between cross-sectional and panel data before and after panel consolidation

		Gender (% female)								
		17	18	19	20	22	24	26	28	30
Data		t5	t6	t7	t8	t9	t10	t11	t12	t13
CS		50.2	53.0	53.0	53.2	53.2	54.3	54.5	54.3	54.1
oPCS		54.3	56.8	56.6	57.8	59.1	60.9	58.6	61.8	62.3
oDiff	%	4.1	3.8	6.6	4.6	5.9	5.6	4.1	6.5	8.2
cPCS		54.2	56.4	54.9	54.8	56.0	57.5	58.1	57.1	58.0
cDiff.	%	4.0	3.4	4.9	1.6	2.8	3.2	3.6	2.8	3.9

oDiff. %= $\%oPCS_t$;- $\%CS_t$; cDiff %= $\%cPCS_t$ - $\%CS_t$

Improving the number of cases of relevant subgroups: The development of parents in the CrimoC-data is displayed in table 8. The number from the respective cross-section data serves as the reference category. The number of parents from the original panel and the consolidated panel are compared with this. The criterion is considered fulfilled if the proportion of parents in the consolidated data set is higher than that of the original data set.

Table 8 Development cases parents between cross-sectional and panel data before and after panel consolidation

		Number of parents				
Data		22 t9	24 t10	26 t11	28 t12	30 t13
CS	n	168	286	490	732	1.004
oPCS	n	106	153	260	392	540
% CS		63.1	53.5	53.1	53.6	53.8
cPCS	n	154	214	386	590	823
% CS		91.7	74.8	78.8	80.6	82.0

% CS= Percentage of cases in relation to the cross-section.

The original panel data set includes only about half of the parents from the cross-sectional data at four of the five points in time shown in the table. Through panel consolidation, the proportion of parents could be drastically increased to 75-82%. In figures, this means, for example, that in t11 126 parents could be subsequently matched, in t13 even 283. Criterion 3 is thus fulfilled.

No changes in the structure of the dependent variable: In the present criminological study, the extent of delinquent behavior is of particular importance. This can be operationalized in two different ways per survey time: A sum index of the annual prevalence rates over the queried 15 offenses (Have you committed the offense in the last 12 months?). This so-called *versatility score* thus has a range of values from 0 to 15. 0 means that an individual has committed none of the offenses, 15 means that an individual has committed all of the offenses queried, while the values in between indicate the respective number of types of offense committed. Strictly speaking, this score measures the number of different types of offense committed. The second possibility is a sum index of the *incidence rates* for each survey time. The incidence corresponds to the frequency of offenses committed within the last 12 months.

However, this sum score is very susceptible to extreme values. For this reason, criminology usually uses the versatility score for complex models, which has proven to be a comparable, less distributionally skewed alternative to the incidence rates (Sweeten, 2012). For both scores, mean values can be found for the different survey waves in table 9. As can be seen, these two variables do not deviate significantly from each other between the two panel data sets, with the mean values of the incidence rates showing somewhat greater deviations than the versatility scores.

Table 9 Versatility scores and incidence rates per time point before and after panel consolidation[5]

Data	Versatility score per time point (and average age)								
	17 t5	18 t6	19 t7	20 t8	22 t9	24 t10	26 t11	28 t12	30 t13
CS_t	0.48	0.27	0.15	0.13	0.10	0.08	0.07	0.06	0.04
oPCS	0.44	0.25	0.14	0.11	0.08	0.06	0.06	0.05	0.04
cPCS	0.44	0.24	0.15	0.12	0.09	0.08	0.06	0.06	0.04
	Incidence rates per time point (and average age)								
CS_t	4.67	4.40	2.50	1.80	0.74	0.57	0.38	0.32	0.31
oPCS	4.82	4.57	2.09	1.52	0.62	0.31	0.28	0.32	0.22
cPCS	4.82	4.51	2.10	1.42	0.66	0.50	0.34	0.33	0.28

Overall, the descriptive results of both panel data sets appear comparable. On the content level, both data sets lead to the same results. Criterion 4 seems to be fulfilled but in longitudinal criminological research, the development of juvenile delinquency is often described using complex trajectories. These are mostly based on *Latent Class Growth Analyses* (LCGA) or on *Growth Mixture Models* (GMM) (Nagin & Land 1993; Vermunt & Magidson, 2004; Muthén, 2004). Using the previously reported versatility score, LCGAs will be calculated for the original and the consolidated panel for two different age periods. Missing values were accounted for using the *full information maximum likelihood estimator* (FIML). In order to check the comparability of both data sets (original versus consolidated panel), two LCGAs are calculated. The first covers age 13 to 19, thus also including the first four waves that were not affected by the consolidation. All cases with a maximum of one missing participation were included in this analysis (original n= 1,907; consolidated n= 2,051). Since the description of the consolidation could show that more missing units could be filled with data at later points in time, another model will be calculated for age 20-30 and up to two missing participations will be tolerated here (original n= 1,865; consolidated n= 2,419). Since the comparability of the results is the focus of this paper, the detailed description of the modelling is omitted (the necessary information can be found in appendix D). Instead, the class solutions found for the original and the consolidated panel are cross-tabulated. The

5 The tables are always described only from the 5th wave onwards, since the first five waves were already fully matched against each other as part of the original panel construction.

quality criterion is still considered to be fulfilled if the class solutions found for the individual cases do not deviate significantly from each other.

At the beginning, all data sets were tested to determine which distributional assumption best fits the data. Due to the fact that a large number of respondents indicated that they had not committed any crime, the versatility score shows many zeros. It was found that the *negative binomial distribution* assumption best fit the highly right-skewed data. A zero-inflated model did not lead to a substantial improvement of model fit.

For age 13 to 19, both models reach a five-class solution. As expected, the model fit values are higher for the consolidated data set due to the higher number of cases.

The five classes found describe typical developmental patterns of delinquent behavior during youth (table 10). The class of *non-offenders* is characterized by the reporting of no or only very isolated offenses. The *Adolescent limited* class shows higher mean versatility scores in early adolescence but commits fewer and fewer offenses with increasing age. The *early desistance* class shows high delinquency scores at the start of adolescence that steadily decrease with age. Compared to the other groups, the *late onset* group shows its highest delinquency levels later, at age 16. The *persistent* class shows the highest burden of delinquency across all waves, although a decline toward young adulthood is also observed for this group. These patterns are found in both the original and consolidated panel data sets. The proportion of cases attributed to a particular class varies only marginally by a maximum of one percent between the data sets, i.e., the consolidated data set can be considered comparable at the content level even in the case of the LCGA for the juveniles.

Based on the variance and co-variance structure of both data sets the latent classes are estimated quite similar. This is reflected in the fact that in Table 11 the diagonal of the crosstab has the highest numbers. 1,093 of the total of 1,096 non-offenders in the original classification are also assigned to this class in the consolidated data set. In total, only 66 of the original 1,907 cases (=3.4%) were assigned to a different class within the consolidated data set, which indicates a stable class solution.

But what happens in the later waves under the acceptance of more missing units? For this purpose, 1,865 cases of the original panel data set and 2,419 cases of the consolidated panel for the age group 20-30 years were conducted with a maximum of two missing units. Both data sets differ by more than 500 cases.

Table 10 Comparison of the versatility score mean values for each class and age for the original and consolidated panel

	Age						
Class	13	14	15	16	17	18	19
Non-offenders							
Original (57%, n=1,096)	0.07	0.06	0.05	0.04	0.03	0.03	0.02
Consolidated (58%, n=1189)	0.07	0.06	0.05	0.04	0.03	0.03	0.02
Adolescent limited							
Original (15%, n=280)	0.60	0.77	0.57	0.24	0.06	0.01	0.00
Consolidated (12%, n=247)	0.61	0.82	0.59	0.23	0.05	0.01	0.00
Early desistance							
Original (10%, n=198)	2.47	2.46	1.83	1.02	0.42	0.13	0.03
Consolidated (12%, n=241)	2.27	2.29	1.74	1.00	0.43	0.14	0.04
Late onset							
Original (11%, n=213)	0.27	0.56	0.89	1.10	1.03	0.75	0.41
Consolidated (12%, n=241)	0.23	0.50	0.82	1.03	0.98	0.71	0.39
Persistant							
Original (6%, n=120)	3.13	3.74	3.85	3.41	2.60	1.70	0.96
Consolidated (6%, n=133)	3.14	3.81	3.94	3.48	2.63	1.69	0.93

n and % based on the most likely latent class membership

Table 11 Cross-tabulation class solution original and consolidated panel age 13 to 19

	Consolidated classification*					
Original Classification*	Non-offenders	Adolescent limited	Early desistance	Late onset	Persistant	total
Non-offenders	1,093	0	0	3	0	1,096
Adol. limited	24	229	12	15	0	280
Early des.	0	0	197	0	1	198
Late onset	0	0	9	202	2	213
Persistant	0	0	0	0	120	120
Not matched	*72*	*18*	*23*	*21*	*10*	*144*
total	1,189	247	241	241	133	2,051

* n based on the most likely class membership, $\chi^2 = .00068$, $p < .001$

Table 12 Comparison of the versatility score mean values for each class and age for the original and consolidated panel age 20 to 30

	Age					
Class	20	22	24	26	28	30
Non-offenders						
Original (88%, n=1,634)	0.02	0.01	0.01	0.00	0.00	0.00
Consolidated (88%, n=2,130)	0.02	0.01	0.01	0.01	0.01	0.00
Adult onset						
Original (9%, n=165)	0.14	0.16	0.18	0.19	0.20	0.20
Consolidated (8%, n=183)	0.14	0.20	0.25	0.28	0.30	0.28
Late desistance						
Original (2%, n=45)	0.87	0.61	0.30	0.11	0.03	0.00
Consolidated (3%, n=75)	0.93	0.66	0.34	0.12	0.03	0.01
Persistent						
Original (1%, n=21)	1.39	1.52	1.50	1.34	1.09	0.79
Consolidated (1%, n=31)	2.05	1.93	1.70	1.40	1.07	0.77

n and % based on the most likely latent class membership

The class solution (table 12) consists of the *non-offenders*, (individuals who, compared to their peers, do not start committing offenses until adulthood (*adult onset*), individuals who do not stop committing offenses in adolescence but in young adulthood (*late desistance*)), and the *persistent offenders*, who commit a comparatively large number of offenses even in adulthood. The percentages of participants in the groups are comparable. Overall, less delinquency was reported for this age range.

The final cross-tabulation of both most likely class memberships leads to a stable class solution, as for adolescence (table 13). Only 42 cases of the original classification were assigned to other classes, the number of cases of the diagonal shows the highest values.

Overall, a satisfactory stability and thus comparability of the data sets with respect to the analysis of developmental trajectories can thus be observed.

Table 13 Cross-tabulation class solution original and consolidated panel age 20 to 30

	Consolidated classification*				
Original Classification*	Non-offenders	Adult onset	Late desistance	Persistant	total
Non-offenders	1,634	0	0	0	1,634
Adult onset	35	127	3	0	165
Late des.	0	0	45	0	45
Persistant	0	2	2	17	21
Not matched	*461*	*54*	*25*	*14*	*554*
total	2,130	183	75	31	2,419

* n based on the most likely class membership, $\chi^2 = .00046$, $p<.001$

Discussion

In this paper, the difficulties of missing units in the construction of panel data with self-generated individual codes in the context of anonymous surveys were discussed. Self-generated codes offer the advantage of assuring anonymity to survey participants. At the same time, they have the disadvantage that they only work if the respondents generate the code identically at all times. If no current code of a new case can be assigned to a case in the data set during the panel construction, a missing unit is created. For time and economic reasons, the previous comparisons of the reported 13-wave panel in the past, except for the first four waves, only took place between a current survey and the two previous surveys. It was shown that although this procedure resulted in a usable panel data set, there were still numerous cases that could not previously be assigned to the panel. With the help of so-called panel consolidation, a procedure in which additional comparisons were made with surveys conducted further apart in time, the quality of the previous data was to be increased. Four criteria were used to assess the quality of the consolidated data set: The number of additional cases or the number of reduced missing units, the reduction of socio-demographic bias, improvement of relevant subgroups and stability of the dependent variable (juvenile delinquency).

Panel consolidation allowed 3,589 missing units in the data set to be replaced with empirical data. This is accompanied by a considerable increase in the number of cases in possible subdata sets. This increase is smaller for data sets without acceptance of missing units, but is greater if missing units are also tolerated in the consolidated data set (table 6). It was also shown that the gender bias could be

reduced across all time points (table 7), and that the cases of the subgroup of parents can be increased enormously (table 8; for example, the number of parents in the 2019 panel cross-section could be increased from 540 to 823 (+34.4%)).

In order to be able to classify the analyses carried out so far on the basis of the original panel and their interpretation in comparison to the consolidated data, the central dependent variable was examined as the last criterion for assessing the impact of the consolidation. It was assumed that there was no systematic bias due to the original panel construction if the consolidation data showed comparable results with regard to this variable.

Both, the descriptive analysis and the longitudinal modelling of LCGAs, lead to the result that both data sets do not differ significantly with regard to the outcome for the dependent variable *juvenile delinquency.* However, the panel consolidation could reduce existing biases and optimize the starting point for subgroup analysis.

The limits of self-generated codes are clearly to be named in their susceptibility to error. Some respondents do not answer identically over time, even to questions on selected, time-stable characteristics. Therefore, it is necessary to design the procedure to be error-tolerant.

Overall, however, this code procedure represents a method of guaranteeing anonymity that is comprehensible to participants.

However, this does not mean that panel consolidation was not necessary. Although the process was very time-consuming and personnel-intensive, numerous missing units could be replaced by empirical information. This automatically also means that data imputation techniques can fall back on a more secure basis. Furthermore, panel consolidation helps to increase the number of cases for subgroup analyses.

References

Bentrup, C. (2020a). The Dual Trajectory Approach. Detecting Developmental Behavioural Overlaps in longitudinal and intergenerational research. *Quality & Quantity* 54(1): 43-65. doi: 10.1007/s11135-019-00934-1

Bentrup, C. (2020b). Gewaltsame Erziehung und ihre Folgen im Altersverlauf. *Monatsschrift für Kriminologie und Strafrechtsreform,* 103(2): 97-120. doi: 10.1515/mks-2020-2042

Bentrup, C. (2019). Untersuchungsdesign und Stichproben der Duisburger Kriminalitätsbefragung. In K. Boers, & J. Reinecke (eds.), *Delinquenz im Altersverlauf. Erkenntnisse der Langzeitstudie Kriminalität in der modernen Stadt* (pp. 95-120). Münster, New York: Waxmann.

Bentrup, C. (2018). First Results of Cross-Generational (Dis-)Similarities Between Three CrimoC Generations. The Relationship Between Experienced Violent Parenting Practice, Delinquency and Own Parenting Style. In V. Eichelsheim, & S. van de Weijer (eds.), *Intergenerational Continuity of Criminal and Antisocial Behaviour. An International Overview of Studies (pp. 235-259).* London & New York: Routledge.

Boers, K., & Reinecke, J. (eds.) (2019). *Delinquenz im Altersverlauf. Erkenntnisse der Langzeitstudie Kriminalität in der modernen Stadt*. Münster; New York: Waxmann.

Boers, K., Reinecke, J., Mariotti, L., & Seddig, D. (2010). Explaining the Development of Adolescent Violent Delinquency. *European Journal of Criminology*, 7(6), 499-520. doi: 10.1177/1477370810376572

Daniel, A., & Erdmann, A. (2017). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2013, Zehn-Wellen-Panel. Schriftenreihe: Jugendkriminalität in der modernen Stadt - Methoden Nr. 23,* Münster, Bielefeld.

Erdmann, A. (2021). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2019 – Dreizehn-Wellen-Panel. Schriftenreihe Kriminalität in der modernen Stadt – Methoden, Heft 27.* Münster, Bielefeld.

Kleinke, K.; Reinecke, J.; Salfrán, D., & Spiess, M. (2020). *Applied Multiple Imputation. Advantages, Pitfalls, New Developments and Applications in R*. Wiesbaden: Springer VS.

Kleinke, K., Reinecke, J., & Weins, C. (2021). The Development of Delinquency During Adolescence: A Comparison of Missing Data Techniques Revisited. *Quality & Quantity* 55(3), 877-895. doi: 10.1007/s11135-020-01030-5

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of Quantitative Methodology for the Social Sciences (pp. 345–368).* Newbury Park, CA: Sage Publications.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology,* 31, 327–362. doi: 10.1111/j.1745-9125.1993.tb01133.x

Pöge, A. (2008). Persönliche Codes „reloaded“. *Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung*, 2 (1), 59-70.

Pöge, A. (2007). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2005, Vier-Wellen-Panel.* Schriftenreihe: Jugendkriminalität in der modernen Stadt- Methoden Nr. 13. Münster, Bielefeld.

Pöge, A. (2005). Persönliche Codes bei Längsschnittstudien. Ein Erfahrungsbericht. *ZA-Information*, 56, 50-69.

Reinecke, J.; Meyer, M., & Boers, K. (2015). Stage-Sequential Growth Mixture Modeling of Criminological Panel Data. In M. Stemmler, A. von Eye, & W. Wiedermann (eds.), *Dependent Data in Social Science Research (pp. 67-89).* Wiesbaden: Springer VS.

Reinecke, J., & Weins, C. (2013). The development of delinquency during adolescence: a comparison of missing data techniques. *Quality & Quantity*, *47*(6), 3319-3334. doi: 10.1007/s11135-012-9721-4

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Seddig, D., & Reinecke, J. (2017). Exploration and Explanation of Adolescent Self-Reported Delinquency Trajectories in the Crimoc Study. In A. Blokland, & V. van der Geest (eds.), *The Routledge International Handbook of Life-Course Criminology (pp. 159-178).* London: Taylor & Francis.

Sweeten, G. (2012). Scaling criminal offending. *Journal of Quantitative Criminology*, *28*(3), 533-557. doi: 10.1007/s10940-011-9160-8

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 2, 549-553.

Appendix

A The query for creating the individual code

Wenn du eine der Fragen überhaupt nicht beantworten kannst, kreuze bitte kein Feld an!

Hier nun die sechs Fragen zur Erstellung deines persönlichen Codes:

1 Bitte kreuze den **ersten** Buchstaben des Vornamens deines **Vaters** (oder einer Person, die für dich einem Vater am nächsten kommt) an. (z. B. Anton, Bernd, Hans-Peter usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

2 Bitte kreuze den **ersten** Buchstaben des Vornamens deiner **Mutter** (oder einer Person, die für dich einer Mutter am nächsten kommt) an. (z. B. Anna, Beate, Jutta, Maria, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

3 Bitte kreuze den **ersten** Buchstaben deines **Vornamens** an (z. B. Michael, Thomas, Ute usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

4 Bitte kreuze den **Tag** deines **Geburtsdatums** an (z.B. Geburtstag am 7. Januar = 7, am 12. Mai = 12, am 31. Oktober = 31)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

5 Bitte kreuze den **letzten** Buchstaben deiner natürlichen **Haarfarbe** an. (z. B. braun, Glatze, schwarz, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

6 Bitte kreuze den **letzten** Buchstaben deiner **Augenfarbe** an. (z. B. braun, grün, grau, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

Hast du im letzten Jahr an der Befragung teilgenommen? ☐ ja ☐ nein
Hast du im letzten Jahr die Schule gewechselt? ☐ ja ☐ nein
Bist du im letzten Jahr sitzen geblieben? ☐ ja ☐ nein

B All alignments, matches and new case counts of the panel cross-sections

Alignment	Number of checks	New matches	Exhaustion in %	oPCS	cPCS	%
2006 with 2005	169	5	2.96	3,032	3,037	
2006 total	**169**	**5**	**2.96**	**3,032**	**3,037**	**+0.16**
2007 with 2004	290	73	25.17			
2007 with 2003	138	50	36.23			
2007 total	**428**	**123**	**28.74**	**2,587**	**2,710**	**+4.75**
2008 with 2005	349	190	54.44			
2008 with 2004	202	99	49.01			
2008 with 2003	114	44	38.60			
2008 total	**665**	**333**	**50.08**	**2,412**	**2,745**	**+13.81**
2009 with 2006	323	203	62.85			
2009 with 2005	236	125	52.97			
2009 with 2004	158	93	58.86			
2009 with 2003	98	38	38.78			
2009 total	**815**	**459**	**56.32**	**2,304**	**2,763**	**+19.92**
2011 with 2007	300	203	67.67			
2011 with 2006	220	116	52.73			
2011 with 2005	255	149	58.43			
2011 total	**775**	**468**	**60.39**	**1,812**	**2,324**	**+28.26**
2013 with 2008	448	309	68.97			
2013 with 2007	180	95	52.78			
2013 with 2006	170	46	27.06			
2013 with 2005	178	30	16.85			
2013 total	**976**	**480**	**49.12**	**1,912**	**2,392**	**+25.10**
2015 with 2009	524	263	50.19			
2015 with 2008	273	137	50.18			
2015 with 2007	164	68	41.46			
2015 with 2006	154	44	28.57			
2015 total	**1,115**	**512**	**45.92**	**1,812**	**2,324**	**+28.26**
2017 with 2011	488	325	66.60			
2017 with 2009	304	155	50.99			
2017 with 2008	153	69	45.10			
2017 with 2007	113	48	42.48			
2017 total	**1,058**	**597**	**56.43**	**1,760**	**2,357**	**+33.92**
2019 with 2013	489	352	71.98			
2019 with 2011	244	121	49.59			
2019 with 2009	213	95	44.60			
2019 with 2008	121	44	36.36			
2019 total	**1,067**	**612**	**57.36**	**1,670**	**2,282**	**+36.65**
total	**7,068**	**3,589**	**50.78**	**4,076**	**4,076**	**+0.00**

C Two examples of post-hoc matching of units

	Code	Participation last year	Gender	Citizenship	Education	New match
First example of eight new matches over time						
w2	HRS2NU	yes	male	German	low level	
w3	HRS2NU	yes	male	German	low level	
w4	HRS2NU	yes	male	German	low level	
w5	HRS2**D**U	yes	male	German	low level	
w6	HRS2NU	yes	male	German	low level	yes
w7	HRS2**D**U	yes	male	German	low level	yes
w8	HRS2**D**UE	-	male	German	low level	yes
w9	HRS2**D**UE	yes	male	German	low level	yes
w10	HRS2**D**UE	yes	male	German	low level	yes
w11	HRS2N**N**E	yes	male	German	low level	yes
w12	HRS2**B**UW	yes	male	German	low level	yes
w13	HRS2**B**UE	yes	male	German	low level	yes
Second example of one new match						
w2	ENB10NN	yes	male	Turkish	high level	
w3	ENB10NN	yes	male	Turkish	high level	
w4	ENB10NN	yes	male	Turkish	high level	
w5	ENB10NN	yes	male	Turkish	high level	
w6	ENB10NN	yes	male	Turkish	high level	
w7	ENB10NN	yes	male	Turkish	high level	
w8	ENB10NNK	yes	male	Turkish	high level	yes
w9	ENB10NNK	yes	male	Turkish	high level	
w10	ENB10NNK	yes	male	Turkish	high level	
w11	ENB10NNK	yes	male	Turkish	high level	
w12	ENB10NNK	yes	male	Turkish	high level	
w13	ENB10NNK	yes	male	Turkish	high level	

The first example reflects a case that was present from w1 to w5 without missing units in the panel data set before the panel consolidation. It can be seen that up to this point, this case only had an error in the code in w5. During the consolidation process, eight units were added to this individual data set. In all cases the code fit

within the error tolerance and also other visible indicators (such as the similarity of the school name) allowed the conclusion that the newly linked units are the same person. The errors in the code are quite easy to justify. It concerns Co005 (the last letter of the own hair color). Numerous respondents had a problem with the change of the query of the first letter (Co001-Co003) to the last letter. If the respondent now had the hair color "dark brown," the error could be explained with the choice of the first letter. If in addition in w12 and w13 only "brown" was meant by the respondent, this error could also be explained. The second case is an example of an individual data set that had only one missing unit until the panel consolidation, which was closed by the additional matching. In this case, a questionnaire could not be found or it could have been subjectively decided during the handwriting check that the questionnaire from the eighth wave should not be linked to w7 or w6.

D Results LCGAs

Original Panel (age 13-19, n=1,907)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	19,006	19,0840	19,039	1,945.57	0.00
3	18,714	18,814	18,757	290.59	0.03
4	18,541	18,663	18,593	175.57	0.00
5	18,498	18,643	18,560	48.79	0.01
6	18,482	18,649	18,554	23.32	0.21

Consolidated Panel (age 13-19, n=2,051)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	20,599	20,678	20,634	2,041.12	0.00
3	20,262	20,363	20,306	334.21	0.03
4	20,067	20,194	20,124	193,98	0.00
5	20,028	20,174	20,092	48.23	0.01
6	20,000	20,168	20,073	35.22	0.19

Original Panel (age 20-30, n= 1,865)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	4,216	4,288	4,246	479.96	0.00
3	4,164	4,256	4,204	61.69	0.00
4	4,154	4,270	4,203	18.21	0.03
5	4,157	4,296	4,216	4.39	0.47

Consolidated Panel (age 20-30, n= 2,419)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	5,904	5,979	5,938	665.24	0.00
3	5,832	5,930	5,876	77.81	0.00
4	5,811	5,933	5,866	27.84	0.03
5	5,811	5,956	5,876	7.77	0.33