
**Methodendokumentation
der kriminologischen Schülerbefragung in Duisburg -
Panelkonsolidierung**

Christina Beckord, Milena Tzvetkova und Hannah Wittbrodt

**Schriftenreihe:
Jugendkriminalität in der modernen Stadt – Methoden
Nr. 28 / 2022**

ISSN 1610-2819

Informationen zur Schriftenreihe
Jugendkriminalität in der modernen Stadt – Methoden

Herausgeber:

Prof. Dr. Klaus Boers

Institut für Kriminalwissenschaften

Abteilung Kriminologie

Bispinghof 24/25

48 143 Münster

und

Prof. Dr. Jost Reinecke

Fakultät für Soziologie

Universität Bielefeld

Postfach 10 01 31

33 501 Bielefeld

Internet-Adressen

<http://www.jura.uni-muenster.de/kriminologie>

<http://www.uni-bielefeld.de/soz/krimstadt>

<http://www.krimstadt.de>

<http://www.crimoc.org>

ISSN 1610-281

Übersicht der bisherigen Titel der Reihe:

1. Motzke, Katharina / Wittenberg, Jochen (1/2004): Methodendokumentation der kriminologischen Schülerbefragung in Münster 2000.
2. Wittenberg, Jochen (2/2004): Methodendokumentation der kriminologischen Schülerbefragung in Münster 2001.
3. Wittenberg, Jochen / Hilfert, Nicole (3/2004): Methodendokumentation der kriminologischen Schülerbefragung in Bocholt 2001.
4. Wittenberg, Jochen (4/2004): Methodendokumentation der kriminologischen Schülerbefragung in Münster 2002.
5. Motzke, Katharina / Brondies, Marc (5/2004): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002.
6. Brondies, Marc (6/2004): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2003.
7. Wittenberg, Jochen (7/2004): Methodendokumentation der kriminologischen Schülerbefragung in Münster 2003.
8. Brondies, Marc (8/2004): Methodendokumentation der Lehrerbefragung an Münsteraner und Duisburger Schulen 2003. Erhebung durchgeführter Präventionsmaßnahmen.
9. Pöge, Andreas (9/2005): Methodendokumentation der kriminologischen Schülerbefragung in Münster 2000–2003 (Vier-Wellen-Panel).
10. Hilfert, Nicole (10/2005): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2004.
11. Kunadt, Susann (11/2006): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2005.
12. Bentrup, Christina (12/2007): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2006.
13. Pöge, Andreas (13/2007): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2005 (Vier-Wellen-Panel).
14. Pollich, Daniela (14/2007): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 und 2003 (Zwei-Wellen-Panel).

15. Bentrup, Christina (15/2009): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2007.
16. Pollich, Daniela (16/2010): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2007 (Sechs-Wellen-Panel).
17. Bentrup, Christina (17/2010): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2008.
18. Bentrup, Christina (18/2012): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2009.
19. Schulte, Philipp (19/2014): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2009 (Integriertes Hell- und Dunkelfeldpanel).
20. Bentrup, Christina / Verneuer, Lena (20/2014): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2011.
21. Verneuer, Lena (21/2015): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2013.
22. Verneuer, Lena (22/2017): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2015.
23. Daniel, Andreas/ Erdmann, Anke (23/2017): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2013 (Zehn-Wellen-Panel).
24. Bentrup, Christina/ Schmid, Jule/ Tzvetkova, Milena/ Vreden, Carina (24/2018): Codebuch der kriminologischen Schülerbefragung in Münster (2000-2003).
25. Kessler, Georg (25/2019): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2017.
26. Kessler, Georg (26/2019): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2019.
27. Erdmann, Anke (27/2021): Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2019 (Dreizehn-Wellen-Panel).
28. **Bentrup, Christina/ Tzvetkova, Milena/ Wittbrodt, Hannah (28/2022):** Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2019 - Panelkonsolidierung.

Inhaltsverzeichnis

Tabellen- und Abbildungsverzeichnis	vii
1 Einleitung	1
2 Bisheriges Zuordnungsverfahren	3
2.1 Studien- und Codedesign	3
2.2 Probleme bei der Zuordnung	5
2.3 Matching-Verfahren	6
3 Panelkonsolidierung	9
3.1 Panelkonsolidierungsverfahren	9
3.2 Ergebnisse der Panelkonsolidierung	10
4 Zusammenfassung und Ausblick	17
5 Anhang	21

Tabellen- und Abbildungsverzeichnis

Tabellen

3.1	Durchgeführte Kontrollschritte	10
3.2	Anzahl Abgleiche/Matches gesamt	11
3.3	Anzahl Abgleiche/Matches ohne bzw. mit einem Fehler im Code für 2019	12
3.4	Alte/neue Panelquerschnitte	12
3.5	Durchgehende Panel ohne Lücken	13
3.6	Lückenhafte Paneldaten	14

Abbildungen

2.1	Studiendesign der Studie Kriminalität in der modernen Stadt	4
2.2	Matching Verfahren bei der ursprünglichen Panelanbindung	7
5.1	Codegenerierung im Fragebogen 2019	21

1 Einleitung

In den Jahren 2002 bis 2019 wurde in Duisburg die insgesamt 13 Wellen umfassende prospektive Panelstudie *Kriminalität in der modernen Stadt*¹ durchgeführt. Ihr Ziel war es, mithilfe eines Selbstauskunftsfragebogens die Entstehung und Entwicklung abweichender und delinquenter Verhaltensweisen in der Phase der Adoleszenz zu beobachten und zu erklären. Dabei werden neben strukturellen Bedingungen und Prozessen auf der Makroebene als mögliche Erklärungen, auch Einflüsse auf der Meso- und Mikroebene untersucht (z. B. soziale Milieus, moralische Orientierungen, Lebensstil, Freizeitgestaltung, Einstellungen, Normorientierungen und soziales Umfeld; detaillierte Informationen zur Studie können auf der Webseite www.crimoc.org abgerufen werden).

Methodisch zeichnet sich die Studie durch die Verwendung eines kombinierten Panel-Kohorten-Designs aus, bei dem mit Hilfe von Fragebögen über Jahre intra- und interindividuelle Veränderungen analysiert werden konnten. Die Studie begann im Jahr 2002 mit einer Befragung der Grundgesamtheit von 3411 Schülern der 7. Klassenstufe in Duisburg. Die Befragungen wurden bis 2009 jährlich, anschließend alle zwei Jahre durchgeführt. Die Wiederbefragungsquoten lagen dabei im Querschnitt zwischen 85 % und 92 %.² Die letzte Erhebung wurde 2019 durchgeführt, sodass für die Studie insgesamt 13 Erhebungswellen in der Alterspanne vom 13. bis 30. Lebensjahr vorliegen.

Auch bei der Studie *Kriminalität in der modernen Stadt* stellte sich, wie grundsätzlich bei Panelstudien, die Frage, wie die Individualdaten der Teilnehmer³ für die spätere Analyse über den Zeitverlauf miteinander verknüpft werden können. Dabei ist zu berücksichtigen, dass im Rahmen von Panelkonstruktionsverfahren Daten einer neuen Erhebungswelle manchmal keinem früheren Erhebungszeitpunkt zugeordnet werden können – entweder weil eine Person tatsächlich bei einzelnen Wellen nicht teilnimmt (Verweigert), oder weil keine eindeutig passenden Individualdaten aus früheren Wellen ausfindig gemacht werden können.

Um in diesem zweiten Fall fehlende Verknüpfungen zu verhindern, wäre es selbstverständlich am einfachsten, die Klarnamen der Teilnehmer zu verwenden. So könnte eindeutig festgestellt werden, wer an welchen Befragungen teilgenommen hat. Dies hat jedoch neben verschiedenen datenschutzrechtlichen Problemen vor allem den Nachteil,

1 Projektleitung: Prof. Dr. Klaus Boers, Institut für Kriminalwissenschaften, WWU Münster und Prof. Dr. Jost Reinecke, Fakultät für Soziologie, Universität Bielefeld.

2 2003 $n = 3392$; 2004 $n = 3339$; 2005 $n = 3243$; 2006 $n = 4548$; 2007 $n = 3336$; 2008 $n = 3086$; 2009 $n = 3090$; 2011 $n = 3050$; 2013 $n = 2850$; 2015 $n = 2754$; 2017 $n = 2778$; 2019 $n = 2697$.

3 Die in diesem Bericht gewählte männliche Form bezieht sich, sofern nicht anders gekennzeichnet, zugleich auf weibliche, männliche und diverse Personen.

1 Einleitung

dass den Teilnehmern keine Anonymität gewährleistet werden kann, welche gerade bei Kriminalitätsbefragungen ein entscheidender Aspekt für die Teilnahme sowie für ein ehrliches Antwortverhalten ist. Zudem wurden bei den ersten Erhebungswellen Minderjährige befragt (in der ersten Welle 13-Jährige, die im Jahr 2002 eine Duisburger Schule besuchten). Aufgrund des jungen Alters der Teilnehmer mussten ihre Daten besonders geschützt werden. Daher wurden für die Studie *Kriminalität in der modernen Stadt* selbst generierte Personencodes verwendet, welche es ermöglichen, die Einzeldaten über die verschiedenen Wellen hinweg zu verknüpfen und dennoch die Anonymität der Teilnehmer zu gewährleisten. Dies erwies sich über die insgesamt 13 Wellen als überwiegend stabiles Zuordnungsverfahren (Erdmann 2021; Daniel und Erdmann 2017; Pöge 2007). Dennoch gab es bei jeder Erhebung einige Teilnehmer, die an einer Erhebung teilgenommen hatten, dem jeweiligen Paneldatensatz jedoch nicht zugeordnet werden konnten. Zwei der möglichen Ursache stellen die falsche Reproduktion des individuellen Codes und/oder Entscheidungen innerhalb des Zuordnungsverfahrens der Panelkonstruktion dar. Denn innerhalb der Panelkonstruktion wurden zunächst die Codes einer aktuellen Erhebung mit Codes vorheriger Befragungen verglichen, die finale Entscheidung über das Vorliegen eines Matches erfolgte jedoch über einen abschließenden Handschriftenabgleich. Der resultierende Dreizehn-Wellen-Paneldatensatz wies durch diese Ursachen und potenzielle Nichtteilnahmen Lücken, sogenannte *missing units* auf, welche eine Auswertung erschwerten. Grundsätzlich ist es zwar möglich, bei Analysen nur die Fälle einzubeziehen, die keine fehlenden Einheiten aufweisen (*listwise deletion*), dies geht jedoch in der Regel mit einer starken Reduktion der Fallzahl im Datensatz einher und kann, wenn diese Reduktion nicht völlig zufällig erfolgt, zu verzerrten Ergebnissen führen. Insbesondere im Rahmen von Subgruppenanalysen (z.B. Elternteile) zeigte sich eine deutliche Differenz zwischen den Fällen der Querschnittsdaten, die die Daten aller Teilnehmer eines Erhebungszeitpunktes enthielten, und Fällen im Paneldatensatz. Aus diesem Grund wurde versucht, möglichst viele dieser Fälle nachträglich zu verknüpfen. Wie genau diese Anbindung erfolgte und wie erfolgreich die nachträgliche Verknüpfung war, wird auf den folgenden Seiten erläutert.

2 Bisheriges Zuordnungsverfahren

2.1 Studien- und Codedesign

Aufgrund des komplexen Paneldesigns (siehe Abbildung 2.1) wurden durch die Studie *Kriminalität in der modernen Stadt* eine Vielzahl verschiedener Datensätze generiert, welche für unterschiedliche Analysen herangezogen werden können.

Zum besseren Verständnis werden die verschiedenen Begriffe an dieser Stelle kurz vorgestellt. Zunächst zu erwähnen sind die Querschnittsdatsätze für jeden Zeitpunkt t (CS_t). Diese umfassen die Angaben aller Personen, die während einer Erhebungswelle den Fragebogen ausgefüllt, also an dieser Welle teilgenommen haben. Die Panel-Querschnittsdatsätze (PCS_t) beinhalten die gematchten Individualdaten pro Welle, die mit mindestens einem Fall (Individualdatensatz) eines anderen Zeitpunkts verknüpft sind. Vollständige Paneldatsätze sind für Längsschnittuntersuchungen am wertvollsten. Sie enthalten nur die Daten derjenigen Befragten, die im untersuchten Zeitraum teilgenommen haben, und die erfolgreich vollständig mit den vorherigen Individualdaten der interessierenden Zeitperiode verknüpft werden konnten (z.B. P_{t1-t13}). Da nur ein relativ geringer Teil der Befragten an allen Erhebungswellen teilgenommen, bzw. zu allen Wellen zugeordnet werden konnte, werden zudem Paneldatsätze mit fehlenden Einheiten verwendet, die alle Fälle bis auf eine tolerierte Anzahl fehlender Einheiten enthalten ($P_{txi,txj,\dots,tX}$).

Um die Fragebögen aus den verschiedenen Erhebungswellen verknüpfen zu können, wurden individuelle kryptografische Codes verwendet, die über Codeblätter zu Beginn des Fragebogens abgefragt wurden. Jeder Befragte füllte einen Codebogen aus, der zunächst fünf, ab der Erhebung 2003 dann sechs persönliche Fragen enthielt, deren Antwort jeweils einen bestimmten Buchstaben oder eine Zahl repräsentierte und entsprechend zu notieren war. Die Fragen beziehen sich auf unveränderliche Merkmale des Befragten oder seines Umfelds (natürliche Haarfarbe, Vorname des Vaters etc.). Diese Buchstaben-Zahlen-Kombination bildet schließlich den gesamten, für jeden Teilnehmer individuellen Code. In jeder Befragungswelle wurde dieser Code zu Beginn mit den unten dargestellten Fragen erhoben. Da die Codes in jeder Umfrage die gleichen Informationen enthielten, mussten die von derselben Person in den verschiedenen Wellen ausgefüllten Codes identisch sein.

Die Fragen zur Erstellung des Codes beinhalteten:

co001: Der erste Buchstabe des Vornamens des Vaters

co002: Der erste Buchstabe des Vornamens der Mutter

co003: Der erste Buchstabe des Vornamens

2 Bisheriges Zuordnungsverfahren

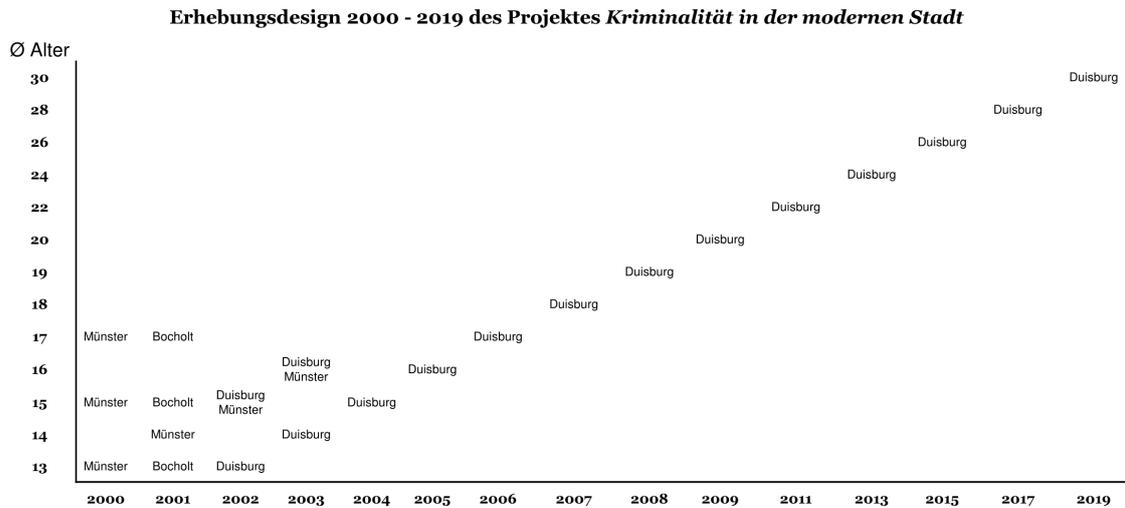


Abbildung 2.1: Studiendesign der Studie *Kriminalität in der modernen Stadt*

co004: Die zweistelligen Ziffern des eigenen Geburtstags

co005: Der letzte Buchstabe der eigenen Haarfarbe

co006: Der letzte Buchstabe der eigenen Augenfarbe

Seit 2009 wird zusätzlich erhoben:

co011: Der letzte Buchstabe des eigenen Nachnamens (bei Namensänderung Geburtsname)

Seit 2003 werden zusätzlich die folgenden Fragen gestellt:

co007: Umfrageteilnahme im Vorjahr (ja/nein)

co008: Schulwechsel im letzten Jahr (ja/nein)

co009: Sitzengeblieben letztes Jahr (ja/nein)

Außerdem wurde das Design geändert: Im Jahr 2002 mussten die Befragten ihre jeweiligen Antworten auf die Codefragen noch handschriftlich in ein Kästchen eintragen. Seit 2003 werden dagegen alle möglichen Buchstaben als Antwortmöglichkeiten zum Ankreuzen angezeigt (siehe Anhang, Seite 21).

Seit der fünften Erhebungswelle (2006) wurde co008 nicht mehr erhoben, da die meisten Befragten ihre Schullaufbahn beendet hatten. Co009 wurde seit 2006 nur noch für diejenigen Befragten erhoben, die ein Gymnasium oder eine Gesamtschule besuchten. Seit der achten Erhebungswelle (2009) wurde nur noch die Erhebungsteilnahme im Vorjahr (co007) abgefragt. Zusätzlich zu den sechs Code-Fragen und den Zusatzfragen standen für die Fragebogenzuordnungen Angaben zum Geschlecht der Befragten und zur (zuletzt) besuchten Schule zur Verfügung.

Für die Funktionalität des Codesystems ist entscheidend, dass durch die Codes eine eindeutige Identifikation der Teilnehmer möglich ist. Dafür muss der Code genügend unterschiedliche Parameter abfragen, welche wiederum eine ausreichende Varianz an Antwortmöglichkeiten aufweisen müssen, sodass die Codes ohne Dopplungen individuell zugeordnet werden können. Des Weiteren müssen die Teilnehmer die Codefragen im Laufe der Zeit auch tatsächlich immer exakt gleich beantworten (Reproduktion). Zudem ist entscheidend, dass es sich bei den abgefragten Merkmalen um zeitinvariante Merkmale handelt, also ausgeschlossen werden kann, dass sich die Antwort im Verlauf der Befragung ändert.

2.2 Probleme bei der Zuordnung

Nach der ersten Befragungswelle im Jahr 2002 fiel jedoch auf, dass mehrfach vollständig gleiche Codes auftraten. Um dieses Identifikationsproblem zu beheben, wurde im folgenden Jahr eine Codefrage hinzugefügt. Dadurch konnte die Eindeutigkeit des Codes stark erhöht werden. Während es im Jahr 2002 noch 324 Dopplungen des vollständigen Codes (7.9 %), 18 Dreifachvorkommen (0.5 %) und 5 Fünffachvorkommen (0.1 %) gab, lag die Anzahl der Dopplungen im Jahr 2003 nur noch bei 32 (0.9 %) (Pöge 2007, S. 6; Pöge 2008, S. 62 f.). Dieser Wert blieb über alle folgenden Erhebungswellen hinweg zwischen 2.0 % im Jahr 2006 und 0.1 % im Jahr 2009. Grundsätzlich waren die Befragten auch mit überwältigender Mehrheit bereit, den Code auszufüllen (zwischen 98.5 % 2005 und 99.6 % 2006).

Es blieb jedoch weiterhin das Problem der Reproduktion. In den ersten vier Wellen lag der Anteil exakt übereinstimmender Codes bei gefundenen passenden Zuordnungen relativ konstant bei lediglich 60 % (berechnet nach Pöge 2007). In den Jahren danach stieg er zunächst etwas an (71 % bei Welle 7) bevor er, vermutlich aufgrund der Umstellung auf den zweijährigen Rhythmus, wieder absank (vgl. Daniel und Erdmann 2017, S. 6). Insbesondere bei den Fragen zu Haar- und Augenfarbe (co005 und co006) fiel es den Teilnehmern offenbar schwer, die Fragen konstant exakt gleich zu beantworten. Dies kann mit dem Wechsel des Frageaufbaus innerhalb des Codes von »erstem Buchstaben« (co001–co003) zu »letztem Buchstaben« (co005 und co006) erklärt werden. Zudem gibt es gerade bei Farben viele verschiedene Antwortmöglichkeiten (und gerade bei Haar- und Augenfarbe auch mögliche Uneindeutigkeiten), sodass nicht unwahrscheinlich ist, dass die Teilnehmer nicht mehr sicher waren, was sie zuletzt angegeben hatten (z. B. Dunkelblond/Hellbraun). Um auch diese Fälle an den Paneldatensatz anbinden zu können, wurde ein fehlertolerantes Matching-Verfahren entwickelt, bei dem schrittweise immer mehr Fehler im Code zugelassen wurden (Pöge 2005a, S. 7 f.; Pöge 2005b, S. 66). Um den Matching-Prozess zusätzlich abzusichern, wurden außerdem alle potenziell passenden Fragebögen einer manuellen Handschriftenkontrolle unterzogen.

2.3 Matching-Verfahren

Dieses fehlertolerante Verfahren mit manuellem Handschriftenvergleich zur Zuordnung der Fragebögen aus den verschiedenen Erhebungsjahren bestand aus vier Hauptschritten: Im ersten Schritt wurden alle exakt übereinstimmenden Codes aus zwei Erhebungswellen identifiziert. Die übereinstimmenden Fragebögen wurden dann einer manuellen Handschriftprüfung nach dem Vier-Augen-Prinzip unterzogen. Dabei wurden auch das Geschlecht und die besuchte Schule abgeglichen. Zusätzlich wurden die Schüler anhand zusätzlicher Variablen (co007, co008, co009) gematcht, welche danach fragten, ob die Schüler im letzten Jahr an der Befragung teilgenommen hatten, sowie, ob sie die Schule gewechselt bzw. nicht gewechselt hatten.⁴

Jene Fragebogenpaare, die offensichtlich von der gleichen Person ausgefüllt wurden, wurden an dieser Stelle aus den zu matchenden Datensätzen entfernt, sodass sie für die nachfolgenden Schritte nicht mehr zur Verfügung standen. Nicht übereinstimmende Fragebögen verblieben in den Datensätzen, um möglicherweise in einem der nächsten Matching-Schritte als übereinstimmend identifiziert zu werden. Im zweiten Schritt wurde unter Zulassung eines Fehlers nach Codematches gesucht, und im dritten Schritt wurden unter Zulassung von zwei Fehlern im Code die zugehörigen Fragebogensnummern ausgeschrieben und manuell überprüft. In einem abschließenden vierten Schritt wurden die Kontrollen unter Zulassung von drei Fehlern im Code durchgeführt. Zudem wurden in jedem Kontrollschritt – unter Beibehaltung der vorgegebenen Fehlerzahl im Personencode – die zusätzlich zu erfüllenden Bedingungen (Geschlecht, besuchte Schule, Teilnahme im Vorjahr) sukzessive gelockert und teilweise ganz weggelassen. Das beschriebene Verfahren ist also hierarchisch und codegeleitet: Das entscheidende Kriterium für eine Zuordnung ist der Code, schrittweise werden Fehler zugelassen und nur offensichtliche Fehlzuordnungen aussortiert. Außerdem wird durch die schrittweise Umsetzung sichergestellt, dass Zuordnungsmöglichkeiten mit weniger Fehlern im Code bevorzugt werden. Die Ergebnisse zeigen, dass, wie zu erwarten, zwar die Zahl der potentiellen Matches stieg, je mehr Fehler zugelassen wurden, jedoch die Anzahl der tatsächlichen Matches sank. Die ersten vier Erhebungswellen wurden mit Hilfe dieses Verfahrens vollständig miteinander abgeglichen. Da es jedoch mit jeder weiteren Welle Abgleiche mit einer größeren Anzahl vorheriger Wellen möglich gewesen wären, was eine Erhöhung des Kosten- und Zeitaufwands bedeutet hätte, wurden aus ökonomischen Gründen ab der fünften Erhebungswelle im Jahr 2006 die Bögen eines Querschnittsdatsatzes, die nach dem oben beschriebenen Matching mit der unmittelbar vorangegangenen Welle noch nicht dem Panel zugeordnet werden konnten, nur noch mit den nicht zugeordneten Bögen der vorletzten Welle abgeglichen.

So sollte sichergestellt werden, dass nicht dauerhaft Personen aus dem Panel fallen, die nur in einem Erhebungsjahr nicht teilgenommen hatten. Dieses Verfahren wurde so

4 Für weitere Informationen zum Matching Prozess vgl. Daniel und Erdmann 2017, S. 8 ff.

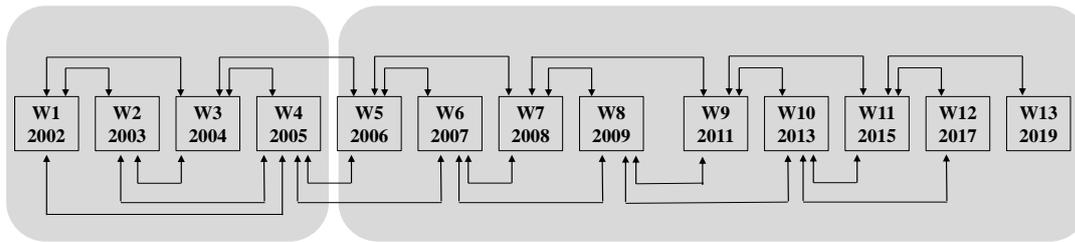


Abbildung 2.2: Matching Verfahren bei der ursprünglichen Panelanbindung

bis 2015 weitergeführt. Ein Matching-Prozess fand also seit 2006 immer nur mit den Erhebungen der letzten beiden Wellen statt (siehe Abbildung 2.2).

Es zeigte sich, dass in den ersten vier Erhebungswellen deutlich mehr Teilnehmer zugeordnet werden konnten als in den Jahren danach. Trotz des Abgleichs mit den beiden vorherigen Wellen konnten seit der sechsten Erhebungswelle im Jahr 2006 zwischen 21.8 % und 38.1 % der teilnehmenden Personen auch durch dieses Vorgehen nicht dem Panel zugeordnet werden. Dadurch verloren der Paneldatensatz und die Panelquerschnitte gegenüber den einfachen Querschnitten eine Vielzahl an Daten, obwohl theoretisch eine Anbindung möglich sein müsste. Mit anderen Worten war zu vermuten, dass der Paneldatensatz *missing units* enthielt, obwohl von einem Fall Daten auf Querschnittsebene vorlagen.

3 Panelkonsolidierung

3.1 Panelkonsolidierungsverfahren

Den sogenannten *missing units* wurde im Paneldatensatz zunächst mit multiplen Imputationsverfahren begegnet. Da diese jedoch nicht in allen Fällen Sinn ergeben, wurde im Jahr 2015 versucht, weitere Personen an den Paneldatensatz anzubinden. Dafür wurden (zunächst nur für die Subgruppe der Elternteile) zusätzliche Codeabgleiche, ähnlich denen des ursprünglichen Matchingverfahrens, für Erhebungszeiträume durchgeführt, die mehr als zwei Wellen auseinanderliegen. Dieses Verfahren stellte sich als überraschend erfolgreich heraus und wurde daraufhin auf das vollständige Panel angewendet. Die sogenannte Panelkonsolidierung sollte dazu verhelfen, Lücken in Form von *missing units* mit empirischen Informationen zu füllen und so die Informationen auf Individualebene zu vervollständigen. Ihr Ziel war es nicht, dem Datensatz neue Fälle in Form von Angaben bisher nicht berücksichtigter Personen anzubinden. Die Fallzahl des ursprünglichen Paneldatensatzes von 4076 Fällen sollte sich somit nicht ändern.

Die Vorgehensweise verlief analog zur ursprünglichen Panelanbindung: Zunächst wurden die Fälle des Querschnitts eines Jahres ausgewählt, die bisher nicht dem Panel zugeordnet werden konnten. Anschließend wurden alle Fälle entfernt, die bereits mit dem ausgewählten Jahr verknüpft waren. Die beiden daraus resultierenden Teildatensätze wurden anschließend auf identische Codes, den potentielle Matches, unter Berücksichtigung des fehlertoleranten Verfahrens untersucht. Dies geschah zunächst mit Hilfe von Access SQL-Abfragen, später wurden die entsprechenden Listen mit der Statistiksoftware R erstellt.⁵

Da diese Teilstichproben deutlich kleiner waren als bei den in der Vergangenheit erfolgten Panelprüfungen, wurde das Matching nur in zwei Schritten durchgeführt (siehe Tabelle 3.1): Schritt 1 umfasste alle Fälle mit identischen Codes für jedes Match, wobei die zusätzlichen Variablen (Geschlecht, Schule, Teilnahme im Vorjahr) nicht übereinstimmen mussten. Im zweiten Schritt wurde ein Fehler im Code toleriert, die weiteren Variablen blieben auch hier unberücksichtigt.

Die durch dieses Verfahren entstandenen potentiellen Matches wurden anschließend manuell von zwei Personen (Vier-Augen-Prinzip) bestätigt oder verneint. Dabei wurden die beiden aufgrund ihres Codes zueinanderpassenden Fragebögen in Papierform auf Plausibilität der Übereinstimmung überprüft. Dies erfolgte zunächst durch eine Handschriftenkontrolle bei den Fragebögen. Konnte durch diese noch keine eindeutige Aussage

⁵ Eine Beispieldatei der R-Syntax findet sich im Anhang auf Seite 22.

3 Panelkonsolidierung

Tabelle 3.1: Durchgeführte Kontrollschritte

Schritt	Code Variablen	Zusatzvariablen
S1	Ohne Fehler	Keine Bedingung
S2	Ein Fehler	Keine Bedingung

getroffen werden, folgte zudem eine Kontrolle von typischerweise zeitinvarianten Angaben in den Fragebögen selbst (beispielsweise Geburtsland oder Geburtsland der Eltern). Nur wenn dadurch eindeutig bestätigt werden konnte, dass es sich um Fragebögen der selben Person handelte, wurden die Fragebögen gematcht. Auf Kontrollen mit mehr als einem Fehler im Code wurde an dieser Stelle, anders als bei der ursprünglichen Panelanbindung, bewusst verzichtet. Grund dafür ist der manuelle Handschriftenvergleich, der vor allem bei Zuordnungen mit mehr als einem tolerierten Fehler an Bedeutung gewinnt, aber mit größer werdenden zeitlichen Abständen auch zunehmend schwieriger wird. Zudem zeigte sich, dass bei mehr als vier weiteren zurückliegenden Zeitpunkten die Zuordnungswahrscheinlichkeit sehr gering wurde und somit auf Abgleiche mit einem zeitlichen Abstand von mehr als vier Erhebungen verzichtet wurde.

3.2 Ergebnisse der Panelkonsolidierung

Insgesamt wurden im Rahmen dieser zweiten Panelanbindung 7164 Abgleiche durchgeführt, von denen 3589 (50.1 %) zu einer neuen Verknüpfung im bestehenden Paneldatensatz führten. Dies ist eine beeindruckende Zahl, wenn man bedenkt, dass zuvor bereits eine Anbindung mit den beiden vorherigen Wellen stattfand.

Tabelle 3.2 zeigt die Anzahl der Abgleiche und Matches in den jeweiligen Jahren. Begonnen wurde die Panelkonsolidierung mit dem Abgleich der noch nicht angebotenen 2006er Fälle. Da, wie bereits erläutert, die erhobenen Daten der ersten vier Wellen (2002-2005) bereits im Rahmen der ursprünglichen Panelkonstruktion vollständig miteinander abgeglichen worden waren, ist es nicht verwunderlich, dass bei einem Abgleich der Daten des Jahres 2006 mit den Daten des Jahres 2003 nur 160 Abgleiche, aus denen fünf neue Zuordnungen hervorgingen, durchgeführt wurden. Die Anzahl an getätigten Abgleichen nahm kontinuierlich zu, je weiter die jeweils interessierende Welle von dem ursprünglich vollständig kontrollierten Zeitraum 2002-2005 entfernt lag. Die meisten Neuzuordnungen konnten schlussendlich für die letzte Erhebungswelle verzeichnet werden. Die meisten Anbindungen gelangen bei den Abgleichen 2019/2013 (352 Matches), 2017/2011 (325 Matches) und 2013/2008 (309 Matches).

Insgesamt konnten weitere 612 Fälle von der Erhebungswelle 2019 an den Paneldatensatz angebotnen werden. Bei der Welle 2017 gelang die Anbindung von 597 Fällen, 2015 waren es 512 Fälle, 2013 480 Fälle, 2011 468 Fälle, 2009 459 Fälle, 2008 333 Fälle, 2007

3.2 Ergebnisse der Panelkonsolidierung

Tabelle 3.2: Anzahl Abgleiche/Matches gesamt

Erhebung	Verwendete Welle zur Panelkonstruktion								
	2013	2011	2009	2008	2007	2006	2005	2004	2003
2019	489/352 (72.0 %)	244/121 (49.6 %)	213/95 (44.6 %)	121/44 (36.4 %)	x x	x x	x x	x x	x x
2017	x x	488/325 (66.6 %)	304/155 (51.0 %)	153/69 (45.1 %)	113/48 (42.5 %)	x x	x x	x x	x x
2015	x x	x x	524/263 (50.2 %)	273/137 (50.2 %)	164/68 (41.5 %)	154/44 (28.6 %)	x x	x x	x x
2013	x x	x x	x x	448/309 (69.0 %)	180/95 (52.8 %)	170/46 (27.1 %)	178/30 (16.9 %)	x x	x x
2011	x x	x x	x x	x x	319/203 (63.6 %)	234/116 (49.6 %)	255/149 (58.4 %)	x x	x x
2009	x x	x x	x x	x x	x x	323/203 (62.8 %)	236/125 (53.0 %)	158/93 (58.9 %)	98/38 (38.8 %)
2008	x x	x x	x x	x x	x x	x x	349/190 (54.4 %)	202/99 (49.0 %)	114/44 (38.6 %)
2007	x x	x x	x x	x x	x x	x x	x x	290/73 (25.2 %)	201/50 (24.9 %)
2006	x x	x x	x x	x x	x x	x x	x x	x x	169/5 (3.0 %)

Prozentangabe in Klammern: Anteil der Matches gemessen an den Abgleichen

immerhin noch 123 Fälle und 2006 5 Fälle. Es ist wichtig zu beachten, dass es nicht Zweck der Panelkonsolidierung war, den Datensatz um neue Fälle zu ergänzen, sondern darum, Lücken (in Form fehlender Einheiten) durch nachträgliche Prüfungen zu schließen. Die Gesamtzahl der Fälle vor und nach der Panelkonsolidierung ist mit je 4076 Fällen folglich identisch.

Um das genaue Vorgehen zu verdeutlichen sind in Tabelle 3.3 beispielhaft die erfolgten Abgleiche und Zuordnungen für das Jahr 2019 für die beiden Schritte der Konsolidierung zusammengefasst. Es zeigt sich erwartungskonform, dass die Abgleiche, die auf Grundlage von identischen Codes für die jeweils betrachteten Zeitpunkte getätigt wurden, zu einer höheren Zuordnungsquote führten als die Abgleiche mit von potentiellen Matches mit einem Fehler im Code.

Vergleicht man die Panelquerschnitte vor und nach der Panelkonsolidierung (Tabelle 3.4), ist die Bedeutung der Konsolidierung deutlich zu erkennen. Zwar gibt es immer noch erkennbare Unterschiede zwischen den Fallzahlen in den Querschnitten und den

3 Panelkonsolidierung

Tabelle 3.3: Anzahl Abgleiche/Matches ohne bzw. mit einem Fehler im Code für 2019

Erhebung 2019	Verwendete Welle zur Panelkonstruktion			
	2013	2011	2009	2008
Abgleiche/Matches ohne Fehler im Code	316/256 (81.0 %)	133/86 (64.7 %)	108/59 (54.6 %)	50/21 (42.0 %)
Abgleiche/Matches mit einem Fehler im Code	173/96 (55.5 %)	111/35 (31.5 %)	105/36 (34.3 %)	71/23 (32.4 %)

Prozentangabe in Klammern: Anteil der Matches gemessen an den Abgleichen

Fallzahlen im Panelquerschnitt, es konnten jedoch bis zu 89.4 % der Fälle (im Jahr 2009) zugeordnet werden, was einen Anstieg um 14.8 % darstellt. Die Tabelle zeigt deutlich, dass insbesondere in den letzten Erhebungswellen (2011–2019) durch die Panelkonsolidierung deutlich mehr Fälle an den Paneldatensatz angebunden werden konnten. Während nach der ersten Anbindung nur zwischen 60 % und 70 % der Querschnittfälle auch dem Panel zugeordnet werden konnten, lagen die Zuordnungswerte nach der Panelkonsolidierung bei 84 %.⁶ Dadurch konnten viele Lücken im Paneldatensatz und somit auch in den Panelquerschnittsdatsätzen, vor allem in den späteren Erhebungswellen, geschlossen werden.

Tabelle 3.4: Alte/neue Panelquerschnitte

	2006	2007	2008	2009	2011	2013	2015	2017	2019
Querschnitt	4548	3336	3086	3090	3050	2849	2754	2778	2698
Panelquerschnitt alt	3032	2587	2412	2304	2100	1912	1812	1760	1671
Differenz alt	1516	749	674	786	950	937	942	1020	1027
Zugeordnet alt in %	66.7 %	77.5 %	78.2 %	74.6 %	68.9 %	67.1 %	65.8 %	63.4 %	61.9 %
Matches Panelkonsolidierung	5	123	333	459	468	480	512	597	612
Panelquerschnitt neu	3037	2710	2745	2763	2568	2392	2324	2357	2283
Differenz neu absolut	1511	626	341	327	515	457	430	421	415
Zugeordnet neu in %	66.8 %	81.2 %	89.0 %	89.4 %	84.2 %	84.0 %	84.4 %	84.8 %	84.6 %
Anstieg Panelquerschnitt	0.1 %	3.7 %	10.8 %	14.8 %	15.3 %	16.9 %	18.6 %	21.4 %	22.7 %

⁶ Die geringe Anbindungsquote im Jahr 2006 lässt sich auf eine Erhöhung des Stichprobenumfangs im Jahr 2006 zurückführen. Um möglichst viele der bisherigen Teilnehmer zu erreichen, wurde die fünfte Erhebung im Jahr 2006 ebenfalls an Duisburger Berufsschulen durchgeführt. Dies führte zu einer erheblichen Vergrößerung des Querschnittsdatsatzes (vgl. Daniel und Erdmann 2017, S. 2). Da die Befragung auch an den Berufsschulen im Klassenkontext durchgeführt werden musste, wurden zunächst viele Personen befragt, die nicht der ursprünglichen Kohorte angehörten. Diese Fälle sind im Querschnittsdatsatz enthalten, sollen, konnten und können jedoch nicht dem Paneldatensatz zugeordnet werden.

3.2 Ergebnisse der Panelkonsolidierung

Tabelle 3.5: Durchgehende Panel ohne Lücken

Anzahl Wellen	Fallzahl alt	Fallzahl neu	Differenz absolut	Differenz in %
13-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8,9,10,11,12,13}$	736	814	+78	+10.6 %
12-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8,9,10,11,12}$	790	874	+84	+10.6 %
$P_{2,3,4,5,6,7,8,9,10,11,12,13}$	884	979	+95	+10.7 %
11-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8,9,10,11}$	832	926	+94	+11.3 %
$P_{2,3,4,5,6,7,8,9,10,11,12}$	950	1057	+107	+11.3 %
$P_{3,4,5,6,7,8,9,10,11,12,13}$	1022	1134	+112	+11.0 %
10-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8,9,10}$	911	1023	+112	+12.3 %
$P_{2,3,4,5,6,7,8,9,10,11}$	1000	1120	+120	+12.0 %
$P_{3,4,5,6,7,8,9,10,11,12}$	1100	1226	+126	+11.5 %
$P_{4,5,6,7,8,9,10,11,12,13}$	1127	1252	+125	+11.1 %
9-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8,9}$	1002	1116	+114	+11.4 %
$P_{2,3,4,5,6,7,8,9,10}$	1098	1242	+144	+13.1 %
$P_{3,4,5,6,7,8,9,10,11}$	1163	1303	+140	+12.0 %
$P_{4,5,6,7,8,9,10,11,12}$	1214	1353	+139	+12.4 %
$P_{5,6,7,8,9,10,11,12,13}$	1246	1356	+110	+10.4 %
8-Wellen-Panel				
$P_{1,2,3,4,5,6,7,8}$	1103	1192	+89	+8.1 %
$P_{2,3,4,5,6,7,8,9}$	1222	1362	+140	+11.5 %
$P_{3,4,5,6,7,8,9,10}$	1274	1439	+165	+13.0 %
$P_{4,5,6,7,8,9,10,11}$	1285	1439	+154	+12.0 %
$P_{5,6,7,8,9,10,11,12}$	1339	1485	+146	+10.9 %
$P_{6,7,8,9,10,11,12,13}$	1294	1438	+144	+11.1 %
7-Wellen-Panel				
$P_{1,2,3,4,5,6,7}$	1189	1267	+78	+6.6 %
$P_{2,3,4,5,6,7,8}$	1365	1468	+103	+7.5 %
$P_{3,4,5,6,7,8,9}$	1424	1589	+165	+11.6 %
$P_{4,5,6,7,8,9,10}$	1415	1593	+178	+12.6 %
$P_{5,6,7,8,9,10,11}$	1419	1581	+162	+11.4 %
$P_{6,7,8,9,10,11,12}$	1394	1553	+159	+11.4 %
$P_{7,8,9,10,11,12,13}$	1327	1498	+171	+12.9 %

Tabelle 3.6: Lückenhafte Paneldaten

Anzahl Lücken	Fallzahl alt	Fallzahl neu	Differenz absolut	Differenz in %	kH alt	kH alt in %	kH neu	kH neu in %	Differenz kH
0	736	814	+79	+10.7 %	736	18.0 %	814	20.0 %	+78
1	494	590	+95	+19.2 %	1230	30.2 %	1404	34.4 %	+174
2	312	430	+118	+37.9 %	1542	37.8 %	1834	45.0 %	+292
3	207	327	+120	+58.0 %	1749	42.9 %	2161	53.0 %	+412
4	216	305	+89	+41.2 %	1965	48.2 %	2466	60.5 %	+501
5	178	181	+3	+1.7 %	2143	52.6 %	2647	64.9 %	+504
6	173	188	+15	+8.7 %	2316	56.8 %	2835	69.6 %	+519
7	181	148	-33	-18.2 %	2497	61.3 %	2983	73.2 %	+486
8	318	162	-156	-49.1 %	2815	69.0 %	3145	77.2 %	+330
9	348	231	-117	-33.6 %	3163	77.6 %	3376	82.8 %	+213
10	387	253	-134	-34.6 %	3550	87.1 %	3629	89.0 %	+79
11	512	434	-78	-15.2 %	4062	99.7 %	4063	99.7 %	+1
12*	14	13	-1	-7.1 %	4076	100.0 %	4076	100.0 %	0

kH: kumulierte Häufigkeiten

*In diesen Fällen wurden Datensätze zwar für mindestens zwei Zeitpunkte einander zugeordnet, jedoch erwies sich ein Fall als qualitativ nicht verwertbar.

Dies zeigt sich nicht nur in den Panelquerschnitten. Wirft man einen Blick auf die gesamten Paneldaten (Tabellen 3.4, 3.5 und 3.6) zeigt sich hier deutlich, wie viele Lücken durch die nachträgliche Einbindung empirischer Informationen geschlossen werden konnten. Alle berichteten durchgehenden Teipaneldatensätze umfassen nun deutlich mehr Fälle (Tabelle 3.5). Insbesondere fallen die 78 neuen Fälle im lückenlosen 13-Wellen-Paneldatensatz auf. Das bedeutet es gibt nun 814 statt 736 Fälle, bei denen eine Person an jeder Erhebungswelle teilgenommen hat. Insbesondere für diese durchgehenden Teildatensätze sind die Ergebnisse überraschend, da bereits bei der normalen Panelanbindung drei aufeinanderfolgende Wellen miteinander verglichen wurden, sodass eigentlich zu erwarten war, dass dabei die Lücken geschlossen worden wären. Erklären lässt sich diese Zahl hauptsächlich dadurch, dass in vielen Fällen mehr als nur eine Lücke innerhalb eines Individualdatensatzes durch die Konsolidierung geschlossen werden konnte. Eine weitere Möglichkeit stellt das Verfahren der Panelkonstruktion an sich dar: Es ist denkbar, dass ein potentielles Match im Rahmen der Handschriftenkontrolle bei der ursprünglichen Konstruktion verneint, bei der späteren Panelkonsolidierung jedoch bejaht wurde.

Zudem beschreibt Tabelle 3.6 detailliert, welche Auswirkungen die Panelkonsolidierung auf Paneldatensätze mit Lücken hat: Die Anzahl der Fälle, die sieben bis zwölf Lücken aufweisen, ist deutlich zurückgegangen. So gibt es z. B. nun 162 Fälle mit acht Lücken im Datensatz, im Vergleich zu 318 Fällen vor der Panelkonsolidierung. Das entspricht einer Abnahme von 49.1 %. Gleichzeitig ist die Anzahl der Fälle, die eine bis sechs Lücken aufweisen, deutlich gestiegen. Die höchste Zunahme verzeichnet das 3-Lücken-Panel,

3.2 Ergebnisse der Panelkonsolidierung

von 207 auf 327 Fälle, diese beträgt entsprechend 58 %. Es kann festgehalten werden, dass vor der Panelkonsolidierung 56.8 % der Fälle keine bis sechs Lücken hatten. Nach der Panelkonsolidierung sind es nun 69.6 %, also ein Anstieg um 519 Fälle. Betrachtet man die Fälle mit null bis sieben Lücken, waren es 61.3 % vor der Panelkonsolidierung und 73.2 % nach der Panelkonsolidierung. Die 70 %-Marke aller Fälle wurde vor der Panelkonsolidierung erstmals beim Addieren aller Fälle mit null bis neun Lücken erreicht (77.6 % der Fälle im Panel wiesen null bis neun Lücken auf). Nach der Panelkonsolidierung ist diese, wie oben gezeigt, mit 73.2 % bereits beim Addieren der Fälle mit null bis sieben Lücken erreicht. Diese positive Entwicklung verbessert die Datenqualität und bietet noch mehr Teilpaneldatensätze mit insgesamt weniger Lücken.

Um die neuen Matches im konsolidierten Paneldatensatz zu markieren, wurden die Variablen `*pk_w#_neu` erstellt (* steht dabei für das Präfix der entsprechenden Welle, also a, b, ..., m; # steht für die Welle, also 1, 2, ..., 13). Diese Variablen nehmen die Werte 1 (neuer Fall im Rahmen der Panelkonsolidierung) und 0 (Fall im ursprünglichen Paneldatensatz vorhanden) an. Mit ihrer Hilfe lassen sich jederzeit die Fälle im konsolidierten Datensatz je nach Bedarf filtern bzw. die neuen Matches schnell identifizieren.

4 Zusammenfassung und Ausblick

Insgesamt lässt sich festhalten, dass die Panelkonsolidierung für die Studie *Kriminalität in der modernen Stadt* sehr erfolgreich war, da durch sie die quantitative Zuordnung real existierender empirischer Informationen und somit die Datenqualität der Paneldatensätze erheblich verbessert werden konnte. Der große zeitliche und auch personelle Aufwand einer erneuten manuellen Kontrolle von 14328 Fragebögen innerhalb von 7164 Abgleichen, hat sich angesichts der 3589 neuen Zuordnungen gelohnt. Es wurden dementsprechend 50.1 % der Abgleiche als Matches kategorisiert, was gleich bedeutend mit der Tatsache ist, dass 3589 *missing units* durch empirische Informationen ersetzt werden konnten. Durch die Anbindung konnten viele Lücken geschlossen werden, was die Arbeit mit den Daten vereinfacht. Dies kann besonders für Untersuchungen der Delinquenz über das Jugendalter hinaus, sowie für Untersuchungen von Delinquenz im jungen Erwachsenenalter oder bestimmte Subgruppenanalysen (z.B. Elternteile). relevant sein.

Die Fallzahl beim lückenlosen 13-Wellen-Panel hat sich um 10.7 % erhöht und beträgt nach der Panelkonsolidierung 814 statt 736. Auch alle Teilpanels weisen nach der Panelkonsolidierung deutlich mehr Fälle auf. Es beinhalten allgemein nun 69.6 % der Fälle null bis sechs Lücken, im Vergleich zu 56.8 % vor der Panelkonsolidierung. Die Zunahme an Fallzahlen vereinfacht es, statistisch signifikante Aussagen zu treffen, sodass zukünftige Analysen mit den Daten der Studie sehr von der Panelkonsolidierung profitieren.

Falls bei der Arbeit mit dem konsolidierten Paneldatensatz eine Filterung der neuen Matches erwünscht sein sollte, kann dies mithilfe der entsprechenden Variablen jederzeit problemlos erreicht werden und somit das Einfügen der neuen Fälle nachverfolgt werden.

Literatur

- Daniel, A. und A. Erdmann (2017). »Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2013, Zehn-Wellen-Panel«. In: Schriftenreihe: Jugendkriminalität in der modernen Stadt – Methoden 23.
- Erdmann, A. (2021). »Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2019, Dreizehn-Wellen-Panel«. In: Schriftenreihe: Jugendkriminalität in der modernen Stadt – Methoden 27.
- Pöge, A. (2005a). »Methodendokumentation der kriminologischen Schülerbefragung in Münster 2000–2003, Vier-Wellen-Panel«. In: *Zeitschrift für Empirische Sozialforschung* 9.
- (2005b). »Persönliche Codes bei Längsschnittstudien. Ein Erfahrungsbericht«. In: *ZA-Information* 56, S. 50–69.
- (2007). »Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2005, Vier-Wellen-Panel«. In: Schriftenreihe: Jugendkriminalität in der modernen Stadt – Methoden 13.
- (2008). »Persönliche Codes ›reloaded‹«. In: *Zeitschrift für Empirische Sozialforschung* 2.1, S. 59–70.

5 Anhang

Liebe Teilnehmerin, lieber Teilnehmer,

da wir Ihren Fragebogen dem der letzten Erhebung ohne Ihren Namen zuordnen wollen, ist es wichtig, dass Sie sich an Ihren persönlichen Code vom vorletzten Jahr erinnern. Denn nur so können Ihre Fragebögen einander zugeordnet werden, ohne dass jemand herausfinden kann, wer diese Fragebögen ausgefüllt hat. Wichtig ist also, dass Sie denselben Code noch wissen. Aus diesem Grund haben wir die nachfolgenden Fragen formuliert, die Ihnen helfen sollen, sich an Ihre persönliche Kombination zu erinnern.

*Bitte kreuzen Sie bei jeder der sieben Fragen immer nur ein Feld an!
Wenn Sie eine der Fragen überhaupt nicht beantworten können, kreuzen Sie bitte kein Feld an!*

Hier nun die sieben Fragen zur Erstellung Ihres persönlichen Codes:

1	<p>Bitte kreuzen Sie den ersten Buchstaben des Vornamens Ihres Vaters (oder einer Person, die für Sie einem Vater am nächsten kommt) an. (z. B. <input type="checkbox"/>Anton, <input type="checkbox"/>Bjerdn, <input type="checkbox"/>Hans-Peter usw.).</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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	<p>Bitte kreuzen Sie den ersten Buchstaben des Vornamens Ihrer Mutter (oder einer Person, die für Sie einer Mutter am nächsten kommt) an. (z. B. <input type="checkbox"/>Anna, <input type="checkbox"/>Beate, <input type="checkbox"/>Jutta, <input type="checkbox"/>Maria, usw.).</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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	<p>Bitte kreuzen Sie den ersten Buchstaben Ihres Vornamens an (z. B. <input type="checkbox"/>Michael, <input type="checkbox"/>Thomas, <input type="checkbox"/>Ute, usw.).</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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	<p>Bitte kreuzen Sie den Tag Ihres Geburtsdatums an (z. B. Geburtstag am 7. Januar = <input type="checkbox"/>7, am 12. Mai = <input type="checkbox"/>12, am 31. Oktober = <input type="checkbox"/>31).</p> <table border="1"> <tr> <td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td><td>13</td><td>14</td><td>15</td> </tr> <tr> <td>16</td><td>17</td><td>18</td><td>19</td><td>20</td><td>21</td><td>22</td><td>23</td><td>24</td><td>25</td><td>26</td><td>27</td><td>28</td><td>29</td><td>30</td><td>31</td> </tr> </table>	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
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	<p>Bitte kreuzen Sie den letzten Buchstaben Ihrer natürlichen Haarfarbe an (z. B. braun<input type="checkbox"/>, Glatz<input type="checkbox"/>, schwarz<input type="checkbox"/>, usw.).</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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	<p>Bitte kreuzen Sie den letzten Buchstaben Ihrer Augenfarbe an (z. B. braun<input type="checkbox"/>, grün<input type="checkbox"/>, grau<input type="checkbox"/>, usw.).</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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																		
7	<p>Bitte kreuzen Sie den letzten Buchstaben Ihres Nachnamens an. (Sollten Sie Ihren Namen gewechselt haben, nehmen Sie Ihren Geburtsnamen!)</p> <table border="1"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> <tr> <td>p</td><td>q</td><td>r</td><td>s</td><td>t</td><td>u</td><td>v</td><td>w</td><td>x</td><td>y</td><td>z</td><td>ä</td><td>ö</td><td>ü</td><td>ß</td> </tr> </table>	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	
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																		

Abbildung 5.1: Codegenerierung im Fragebogen 2019

Listing 5.1: R-Syntax zur Erstellung der Abgleichlisten für die Erhebungsjahre 2019 und 2008

```

1 # Das benoetigte Paket laden , womit die STATA-Datensaetze gelesen werden koennen
2 library (foreign)
3 library (rio)
4 library (haven)
5 library (dplyr)
6
7 # Paneldatensatz einlesen
8 panel <- read.dta (" ... panel_2019_2009_13_wellen_12.dta")
9 dim (panel)
10
11 # Querschnittsdatsatz einlesen
12 quer2019a <- read.dta (" ... kbdu_2019_12.dta")
13 dim (quer2019a)
14
15 # Als ungueltig markierte Faelle aus dem Querschnittsdatsatz herausnehmen
16 quer2019 <- quer2019a [quer2019a$w13boese=="nein" ,]
17 dim (quer2019)
18
19 # Faelle aus 2019, die dem Panel bereits zugeordnet sind
20 in.panel <- panel [panel$welle13==1,]
21 dim (in.panel)
22
23 # Faelle herausfiltern , die fuer 2008 vorhanden , aber nicht an 2019 gematched sind
24 liste2008 <- panel [panel$w7=="vorhanden" & panel$welle13==0,]
25 dim (liste2008)
26
27 # Fragebogennummern aus dem Querschnitt herausfiltern
28 F19 <- c (quer2019$a0000)
29 length (F19)
30
31 # Fragebogennummern der Faelle von 2019, die dem Panel zugeordnet sind
32 F05 <- c (in.panel$ma0000)
33 length (F05)
34
35 # Faelle aus dem Querschnitt herausfiltern , die NICHT dem Panel zugeordnet sind
36 liste2019 <- subset (quer2019 , !(F19 %in% F05))
37 dim (liste2019)
38
39 # Nicht benoetigte Variablen aus den jeweiligen Datensatzen entfernen
40 liste2008sub <- subset (liste2008 , select=c ("ga0000" ,"gco001" ,"gco002" ,
41                                             "gco003" ,"gco004" ,"gco005" ,
42                                             "gco006" ,"ga0011" ,"ga0102"))
43
44 liste2019sub <- subset (liste2019 , select=c ("a0000" ,"co001" ,"co002" ,
45                                             "co003" ,"co004" ,"co005" ,
46                                             "co006" ,"a0011" ,"a0102" ,"modus"))
47
48 # Alle Angaben zum Code auf Grossbuchstaben setzen , da sie sonst als ungleich
   gewertet werden
49 liste2008sub [,2:7] <- mutate_all (liste2008sub [,2:7] , funs (toupper))
50 liste2019sub [,2:7] <- mutate_all (liste2019sub [,2:7] , funs (toupper))
51
52
53 # Liste , die die entsprechenden Fragebogennummern enthaelt fuer Faelle , bei
54 # denen der generierte Code 2008 und 2019 komplett identisch ist
55 liste <- merge (liste2008sub , liste2019sub ,
56                by.x = c ("gco001" ,"gco002" ,"gco003" ,"gco004" ,"gco005" ,"gco006") ,
57                by.y = c ("co001" ,"co002" ,"co003" ,"co004" ,"co005" ,"co006"))
58

```

```

59 # Sortieren der einzelnen Variablen, damit sie uebersichtlicher sind
60 liste_end <- liste[,c(7,10,1,2,3,4,5,6,8,11,9,12,13)]
61
62 liste_end <- liste_end[,c(1,2,13)]
63
64 names(liste_end) <- c("a00_08", "a00_19", "Modus")
65
66 # Abspeichern der fertigen Liste
67 liste_oF <- liste_end[order(liste_end$a00_08),]
68 dim(liste_oF)
69 write.csv(liste_oF, row.names=F, file="...S1_oF.csv")
70
71 #####
72
73 # Faelle, die einen komplett identischen Code aufweisen aus den Datensatzen
  entfernen
74 liste2008sub_1 <- subset(liste2008sub, !(c(liste2008sub$ga0000) %in% c(liste_end$a00_08)))
75 liste2019sub_1 <- subset(liste2019sub, !(c(liste2019sub$a0000) %in% c(liste_end$a00_19)))
76
77 liste_oCo1a <- merge(liste2008sub_1, liste2019sub_1,
78                     by.x = c("gco002", "gco003", "gco004", "gco005", "gco006"),
79                     by.y = c("co002", "co003", "co004", "co005", "co006"))
80
81 liste_oCo1b <- liste_oCo1a[,c(6,10,7,11,1,2,3,4,5,8,12,9,13,14)]
82 liste_oCo1b <- liste_oCo1b[,c(1,2,14)]
83
84 names(liste_oCo1b) <- c("a00_08", "a00_19", "Modus")
85 liste_oCo1 <- liste_oCo1b[order(liste_oCo1b$a00_08),]
86 write.table(liste_oCo1, row.names=F, file="...S2_cCo1_frei.csv")
87
88 #####
89
90 liste_oCo2a <- merge(liste2008sub_1, liste2019sub_1,
91                     by.x = c("gco001", "gco003", "gco004", "gco005", "gco006"),
92                     by.y = c("co001", "co003", "co004", "co005", "co006"))
93
94 liste_oCo2b <- liste_oCo2a[,c(6,10,1,7,11,2,3,4,5,8,12,9,13,14)]
95 liste_oCo2b <- liste_oCo2b[,c(1,2,14)]
96
97 names(liste_oCo2b) <- c("a00_08", "a00_19", "Modus")
98 liste_oCo2 <- liste_oCo2b[order(liste_oCo2b$a00_08),]
99 write.table(liste_oCo2, row.names=F, file="...S2_cCo2_frei.csv")
100
101 #####
102
103 liste_oCo3a <- merge(liste2008sub_1, liste2019sub_1,
104                     by.x = c("gco001", "gco002", "gco004", "gco005", "gco006"),
105                     by.y = c("co001", "co002", "co004", "co005", "co006"))
106
107 liste_oCo3b <- liste_oCo3a[,c(6,10,1,2,7,11,3,4,5,8,12,9,13,14)]
108 liste_oCo3b <- liste_oCo3b[,c(1,2,14)]
109 names(liste_oCo3b) <- c("a00_08", "a00_19", "Modus")
110
111 liste_oCo3 <- liste_oCo3b[order(liste_oCo3b$a00_08),]
112 write.table(liste_oCo3, row.names=F, file="...S2_cCo3_frei.csv")
113
114 #####
115

```

5 Anhang

```
116 liste_oCo4a <- merge(liste2008sub_1, liste2019sub_1,
117                       by.x = c("gco001", "gco002", "gco003", "gco005", "gco006"),
118                       by.y = c("co001", "co002", "co003", "co005", "co006"))
119
120 liste_oCo4b <- liste_oCo4a[,c(6,10,1,2,3,7,11,4,5,8,12,9,13,14)]
121 liste_oCo4b <- liste_oCo4b[,c(1,2,14)]
122
123 names(liste_oCo4b) <- c("a00_08", "a00_19", "Modus")
124 liste_oCo4 <- liste_oCo4b[order(liste_oCo4b$a00_08),]
125 write.table(liste_oCo4, row.names=F, file="...S2_cCo4_frei.csv")
126
127 #####
128
129 liste_oCo5a <- merge(liste2008sub_1, liste2019sub_1,
130                       by.x = c("gco001", "gco002", "gco003", "gco004", "gco006"),
131                       by.y = c("co001", "co002", "co003", "co004", "co006"))
132
133 liste_oCo5b <- liste_oCo5a[,c(6,10,1,2,3,4,7,11,5,8,12,9,13,14)]
134 liste_oCo5b <- liste_oCo5b[,c(1,2,14)]
135
136 names(liste_oCo5b) <- c("a00_08", "a00_19", "Modus")
137 liste_oCo5 <- liste_oCo5b[order(liste_oCo5b$a00_08),]
138 write.table(liste_oCo5, row.names=F, file="...S2_cCo5_frei.csv")
139
140 #####
141
142 liste_oCo6a <- merge(liste2008sub_1, liste2019sub_1,
143                       by.x = c("gco001", "gco002", "gco003", "gco004", "gco005"),
144                       by.y = c("co001", "co002", "co003", "co004", "co005"))
145
146 liste_oCo6b <- liste_oCo6a[,c(6,10,1,2,3,4,5,7,11,8,12,9,13,14)]
147 liste_oCo6b <- liste_oCo6b[,c(1,2,14)]
148
149 names(liste_oCo6b) <- c("a00_08", "a00_19", "Modus")
150 liste_oCo6 <- liste_oCo6b[order(liste_oCo6b$a00_08),]
151 write.table(liste_oCo6, row.names=F, file="...S2_cCo6_frei.csv")
```