
Tobias Brändle, Severin V. Weiland & Rainer Schnell

Perspektiven für ein bundesweites Bildungsverlaufsregister

Eine Analyse des Zentralen Schülerregisters Hamburg

Zusammenfassung

Die Studie untersucht die Machbarkeit eines bundesweiten Bildungsverlaufsregisters anhand des Zentralen Schülerregisters Hamburgs. Durch Record-Linkage-Verfahren auf Basis von Personenmerkmalen kann eine ausreichend hohe Verknüpfungsqualität erreicht werden. Es bestehen jedoch Herausforderungen bei der Verknüpfung von Daten für Migrantinnen und Migranten sowie in Gebieten mit hoher Bevölkerungsdichte. Eine hohe Datenqualität und ein über die Zeit konsistentes ID-Management sind essenziell für die erfolgreiche Implementierung eines Bildungsverlaufsregisters.

Schlüsselwörter: Bildungsverlauf; Linkage; Längsschnittuntersuchung; Register

Perspectives for a Nationwide Educational Trajectory Register

An Analysis of the Central Pupil Register of the City of Hamburg

Abstract

The study examines the feasibility of a nationwide educational trajectory register based on the Central Pupil Register of Hamburg. A sufficiently high linkage quality can be achieved by using record linkage procedures based on personal characteristics. However, there are challenges in linking data for migrants and regions with high population density. High data quality and consistent ID management over time are essential for successfully implementing an educational trajectory register.

Keywords: educational trajectory register; record linkage; data quality; personal characteristics; administrative data; longitudinal data

1 Einleitung

Mit dem Koalitionsvertrag des Bundes (CDU et al., 2025, S. 72) wurde das Bildungsverlaufsregister (BVR) auf die Hauptbühne der bildungspolitischen Debatte verlegt. Dies verleiht einem Vorhaben, das bislang eher ein Nischendasein fristete (Fickermann, 2021), verstärkte Aufmerksamkeit. Eine zentrale Rolle bei der Realisierung dieser Pläne nimmt die Einführung einer Identifikationsnummer (ID) ein: Sie soll zwischen den Ländern kom-

patibel, datenschutzkonform sowie mit der Bürger-ID verknüpfbar sein (CDU et al., 2025, S. 72).

Doch die Diskussion um ein Bildungsverlaufsregister ist nicht neu. Seit einigen Jahren wird die Einführung eines Bildungsverlaufsregisters (BVR) wieder verstärkt diskutiert (Fickermann, 2021). Maßgeblich für diese erneute Fokussierung war ein „Expertenworkshop Bildungsregister“, der im November 2018 vom Bundesministerium für Bildung und Forschung initiiert wurde. Die Gründe für diese erneute Beschäftigung liegen in der Einführung eines registerbasierten Bevölkerungszensus und der Registermodernisierung. Hierin wurde Potenzial für den Aufbau eines bereichsübergreifenden BVRs gesehen, obwohl die Kultusministerkonferenz bereits Anfang der 2000er Jahre mit dem Versuch der Etablierung länderübergreifender Verlaufsdaten auf Individualebene im Schulbereich, vor allem aus Datenschutzgründen, scheiterte (Mundelius, 2019).

Jenseits des Datenschutzes prägen heute die Themen Zuständigkeit, Datenumfang und Datenzugang die Debatte um das BVR. Diese Diskussionen sind geprägt vom Bildungsföderalismus, dem Spannungsfeld zwischen Datenhunger und Datensparsamkeit sowie organisatorischen Fragen zum Datenzugang und der Einrichtung einer Vertrauensstelle (Gawronski, 2020; Giar et al., 2023; Hertweck et al., 2023; RatSWD, 2022). Dennoch sind sich alle Handelnden einig: Ein BVR birgt ein grundlegendes Potenzial zur Verbesserung des deutschen Bildungswesens durch gezielte Datennutzung, erhöhte Transparenz und eine evidenzbasierte Bildungspolitik.

Ebenso besteht – trotz dieser herausfordernden Ausgangslage – Einigkeit über die Voraussetzungen für ein BVR (Kap. 2.2), insbesondere die Notwendigkeit einer qualitativ hochwertigen Datenbasis. Hierfür ist eine ID unerlässlich, um Längsschnittverknüpfungen zu ermöglichen. Wobei es sich bei dieser ID jedoch handelt, ist bislang ungeklärt.

Aus fachlicher Perspektive könnte sie entweder eine Verwaltungs-ID sein oder mittels Record-Linkage-Verfahren aus Personenmerkmalen gebildet werden. Da die Steuer-ID rechtlich nicht als Bildungs-ID nutzbar ist (RatSWD, 2022), ist ein alternatives ID-Management erforderlich (Giar et al., 2023). Ferner gilt es zu beachten, dass eine Verwaltungs-ID nicht vollständig Record-Linkage-Verfahren ersetzen kann. Insbesondere bei der Verknüpfung mit historischen Daten oder Befragungsdaten, bei denen die Verwaltungs-ID nicht vorliegt, benötigt es weiterhin Personenmerkmale für die Verknüpfung.

Eine ID erlaubt die Anonymisierung der Bildungsdaten durch die Trennung von identifizierenden Personenmerkmalen von den für Analysen verfügbaren Merkmalen (Christen et al., 2020). Da derartige Daten jedoch sehr eingeschränkt zugänglich sind, können bislang vorliegende Expertisen und Positionspapiere allerdings keine Aussage zur Qualität der grundlegenden Daten und damit zur Realisierbarkeit eines solchen ID-Managements treffen.

Unsere Auswertungen sollen diese Lücke schließen. Wir nutzen die Hamburgische Schulstatistik für allgemeinbildende Schulen und Daten aus dem Zentralen Schülerregister (ZSR), um zu untersuchen, welche Merkmale erforderlich sind, um Bildungsverläufe unverzerrt abzubilden. Wir können mit diesen Daten, die eine Verwaltungs-ID aufweisen, prüfen, welchen Verknüpfungserfolg wir erreichen, wenn wir nur Personenmerkmale und

nicht die Verwaltungs-ID für die Verknüpfung nutzen. Frühere Studien zeigen, dass Vor- und Nachname, Geburtsdatum, Geschlecht und Geburtsort für eine hohe Verknüpfungsqualität notwendig sind (Schnell, 2022). Auf Grundlage amtlicher Daten analysieren wir, wie die Wahrscheinlichkeit für eine längsschnittliche Rekonstruierbarkeit von Bildungsverläufen nach soziodemografischen Merkmalen variiert und wie die Ausfallwahrscheinlichkeit regional unterschiedlich verteilt ist.

Unsere Analysen zeigen, dass Vor- und Nachname, Geburtsdatum, Geschlecht und eine adressbasierte Gitterkoordinate hinreichende Verknüpfungserfolge ermöglichen. Migranten und Menschen mit Migrationshintergrund haben jedoch eine höhere Wahrscheinlichkeit für Verknüpfungsfehler. Zudem variiert die Ausfallwahrscheinlichkeit regional, selbst in Hamburg. Unsere Ergebnisse tragen zur Diskussion über die Ausgestaltung eines BVRs bei.

2 Das Potenzial eines Bildungsverlaufsregisters

2.1 Hoffnungen

Mit der Einführung eines BVRs sind verschiedene Hoffnungen verbunden. Dazu zählen die Schaffung einer „Grundlage für evidenzbasiertes politisches und administratives Handeln“ (Hertweck et al., 2023, S. 733), die Erhöhung der Transparenz der (Bildungs-)Politik (RatSWD, 2022), die Aufhebung eines Wettbewerbsnachteils für den Forschungsstandort Deutschland durch die Ausweitung der Zugänglichkeit von Bildungsdaten für die Wissenschaft sowie deren Verknüpfbarkeit mit anderen Datenquellen – wie sozialwissenschaftlichen Befragungen (Hertweck et al., 2023; RatSWD, 2022) – oder auch die Evaluation von Investitionen in Bildung (Hertweck et al., 2023; SWK, 2022).

Ein BVR bietet auch individuelle und systemische Vorteile. Wenn Kompetenzdaten in ein Register einfließen, würde es die frühzeitige Identifikation und gezielte Adressierung individueller Förderbedarfe vonseiten der Bildungseinrichtung ermöglichen. Wenn Informationen über die soziale Herkunft erfasst würden, helfe das Register auf systematischer Ebene, Bildungsungleichheiten zu erkennen und Maßnahmen zur Förderung von Bildungsgerechtigkeit zu implementieren. Zudem können langfristige Trends und Entwicklungen im Bildungsbereich besser erkannt werden, was die Bildungsplanung verbessert.

2.2 Voraussetzungen für die Realisierung dieser Potenziale

Die Realisierung all dieser Potenziale ist jedoch an bestimmte Voraussetzungen gebunden. Es fehlt bislang an einer einheitlichen Datengrundlage für Schülerindividualdaten (vgl. Hertweck et al., 2023, S. 735) sowie an entsprechenden gesetzlichen und datenschutzrechtlichen Regelungen (RatSWD, 2022). Auch mangelt es bislang an einer Bildungs-ID, die für bildungsbereichsübergreifende Verknüpfungen notwendig ist (Brändle, 2024; Giar et al., 2023). Eine geeignete technische und organisatorische Infrastruktur ist ebenfalls erforderlich (Giar et al., 2023), einschließlich datenschutzkonformer und interoperabler IT-Lösungen sowie einer Vertrauensstelle. Politische Einigkeit zwischen Bund und Ländern ist hierbei entscheidend.

Zusätzlich zu diesen Voraussetzungen sind bestehende Bedenken gegenüber der Sammlung von Bildungsdaten zu adressieren und auszubalancieren, um Vertrauen in die Sinnhaftigkeit eines BVRs zu schaffen. Insbesondere muss es gelingen, das BVR so auszugestalten, dass dessen Vorteile leicht nachvollziehbar sind und es nicht als Überwachungsinstrument wahrgenommen wird.

2.3 Risiken

Bei der Erfüllung der Voraussetzungen ist besonders darauf zu achten, dass die technische und organisatorische Infrastruktur den Schutz sensibler Daten gewährleistet. Zudem müssen die finanziellen und personellen Ressourcen gesichert sein.

Neben diesen strukturellen Risiken bestehen inhaltliche Risiken bei der Umsetzung eines BVRs. Das BVR muss so ausgestaltet werden, dass die Vielfalt und Dynamik individueller Bildungsverläufe abgebildet werden können, einschließlich nichtlinearer Bildungswege, um eine starre Definition von Bildungserfolg zu vermeiden. Auch muss konzeptionell sichergestellt sein, dass die Bildungsverläufe aller Bildungsteilnehmenden gleichermaßen abgebildet werden können – also das Auftreten systematisch fehlender Werte vermieden wird. Zur Vermeidung dieses Risikos wird ein effizientes und über die Zeit konsistentes ID-Management benötigt.

3 Daten, Variablen und Methode

Das ID-Management ist das Herzstück des BVRs und für die Verknüpfung von Datensätzen unverzichtbar. Diese Verknüpfungen haben das Ziel, Informationen über dieselbe Person zusammenzuführen (auch als Record-Linkage bezeichnet; vgl. Herzog et al., 2007, S. 81). Wenn Daten keine eindeutige Verwaltungs-ID aufweisen, müssen diese Verknüpfungen anhand von nicht eindeutigen und zumeist fehlerbehafteten Personenmerkmalen, wie Name oder Geburtsdatum, vorgenommen werden. Entscheidend für die Qualität der Verknüpfung sind die Eindeutigkeit, zeitliche Stabilität und Fehlerfreiheit der verwendeten Personenmerkmale (Herzog et al., 2007).

Im Folgenden wird anhand von zwei Record-Linkage-Szenarien mit jeweils zwei Record-Linkage-Methoden analysiert, wie Bildungsverlaufsdaten auch ohne Verwaltungs-ID verknüpft werden können. Eine detaillierte Beschreibung der Verknüpfungsverfahren sowie weiterer Methoden und Szenarien findet sich im technischen Bericht (Weiland, 2024).

3.1 Datengrundlage

Die folgenden Analysen basieren auf amtlichen Daten zu Schülerinnen und Schülern an allgemeinbildenden Schulen in Hamburg. Datengrundlage waren das Zentrale Schülerregister (ZSR) sowie der Schülerindividualdatensatz (IVDS; Behörde für Schule und Berufs-

bildung, 2022, 2024).¹ Sie beinhalten eine Verwaltungs-ID, sodass geprüft werden kann, welcher Verknüpfungserfolg erreichbar ist, wenn nur Personenmerkmale und nicht die Verwaltungs-ID für eine Verknüpfung genutzt werden. Diese Personenmerkmale stammen aus dem ZSR. Der IVDS enthält soziodemografische Merkmale und Informationen zum Schulbesuch, die zur Analyse der Verknüpfungsqualität genutzt werden. Hierzu zählt der RISE-Index, der den sozioökonomischen Status eines Hamburger Wohnquartiers beschreibt (Behörde für Stadtentwicklung und Wohnen, 2022).

Die migrationsbezogene Herkunft einer Person wurde aus den Daten des ZSR und IVDS bestimmt, basierend auf Staatsbürgerschaft und Geburtsland. Personen, die in Deutschland geboren sind und ausschließlich die deutsche Staatsbürgerschaft besitzen, wurden als ‚deutsch‘ kodiert. Migranten wurden als Personen definiert, die weder in Deutschland geboren sind noch die deutsche Staatsbürgerschaft besitzen. Alle anderen Fälle wurden als Menschen mit Migrationshintergrund zusammengefasst.²

Zusätzlich wurde OpenStreetMap genutzt, um die geografische Koordinate der im ZSR enthaltenen Wohnorte zu ermitteln. Die Koordinate wurde in eine 100m-Gitterkoordinate überführt (INSPIRE Referenzsystem). Diese bietet aufgrund der höheren Fehlertoleranz gegenüber der exakten Adresse Vorteile für die Datenverknüpfung (Weiland, 2024). Zusätzlich wurden Großgebäude und Großwohnsiedlungen in den Daten betrachtet, da diese aufgrund der großen Anzahl Personen mit derselben Adresse sowie der zumeist höheren Fluktuation Probleme bei einer Verknüpfung verursachen können (Statistische Ämter des Bundes und der Länder, 2004). Großgebäude und Großwohnsiedlungen wurden anhand der Bevölkerungsdichte operationalisiert, also der Anzahl Personen im ZSR, die innerhalb derselben 100m-Gitterzelle leben.

Die Korrektheit der Verknüpfung wird über einen eindeutigen Identifikator geprüft, der in beiden Datenquellen enthalten ist. Mit den verknüpften Datensätzen lassen sich Fragen nach der Eignung amtlicher Daten für den Aufbau eines bundesweiten BVRs näher betrachten. Herangezogen wurden Daten zu Schülerinnen und Schülern, die im Schuljahr 2021/22 ($N = 246.472$) und im Schuljahr 2023/24 ($N = 252.230$) im ZSR – jeweils zu Schuljahresbeginn – geführt wurden.³ Insgesamt 182.385 Schülerinnen und Schüler sind in beiden Datensätzen enthalten und haben daher in beiden Schuljahren in Hamburg eine Schule besucht.

-
- 1 Die Daten aus diesen beiden Quellen wurden nach datenschutzrechtlicher Prüfung zu Analysezwecken durch die Vertrauensstelle der Behörde für Schule, Familie und Berufsbildung zur Verfügung gestellt. Die Verarbeitung erfolgte ausschließlich in einer besonders geschützten Umgebung.
 - 2 Hierbei handelt es sich um eine pragmatische Unterscheidung, die auf Basis der verfügbaren Daten getroffen werden musste.
 - 3 Ursprünglich war angestrebt, die Daten aus zwei aufeinanderfolgenden Schuljahren für das Vorhaben zu nutzen. Infolge des Datums der Antragstellung für die Datenbereitstellung war dies nicht möglich. Die Ziele der Analyse können trotz dieser Datenlücke erreicht werden.

3.2 Szenarien

Die Daten des ZSR und des IVDS wurden in zwei Szenarien verknüpft, wobei jeweils unterschiedliche Personenmerkmale genutzt wurden. Entsprechend dem gegenwärtigen Forschungsstand zu erforderlichen Personenmerkmalen in einem bundesweiten BVR (Schnell, 2022) wurden in Szenario A die Merkmale Vorname, Nachname, Geburtsdatum und Geschlecht verwendet. Der Geburtsort konnte nicht aufgenommen werden, da diese Information nicht verfügbar war. Anzunehmen ist jedoch, dass dieses für ein bundesweites BVR erforderliche Merkmal (Schnell, 2022) aufgrund der geringen Variation innerhalb der Hamburger Schülerschaft keine Auswirkung auf den Verknüpfungserfolg des ZSR hat. Szenario B enthält zusätzlich zu den zuvor genannten Merkmalen eine Adressangabe in Form der 100m-Gitterkoordinate. Deskriptive Angaben über die in den Szenarien verwendeten Personenmerkmale finden sich in Tabelle 1.

Tab. 1: Deskriptive Angaben über die bei der Verknüpfung verwendeten Personenangaben

Merkmal	Szenario A	Szenario B	n _{Ausprägungen}	n _{fehlend}
Vorname	X	X	114.587	750
Nachname	X	X	80.727	28
Geburtstag	X	X	31	0
Geburtsmonat	X	X	12	0
Geburtsjahr	X	X	60	0
Geschlecht	X	X	2	18
100m-Gitter-ID		X	34.163	275

Quelle: eigene Berechnung.

3.3 Datenverknüpfung

Für die Verknüpfung des ZSR werden zwei Record-Linkage-Methoden hinsichtlich des Verknüpfungserfolgs getestet:

- probabilistisches Record-Linkage anhand des Fellegi-Sunter-Modells (Fellegi & Sunter, 1969) und
- deterministisches Record-Linkage anhand von multiplen Matchkeys.

Durch die Verwendung mehrerer Methoden und Szenarien kann untersucht werden, ob bestimmte Subpopulationen schlechter verknüpft werden (Linkage-Bias). Die beiden ausgewählten Methoden zeigten beim Vergleich mit weiteren Methoden die besten Linkage-Ergebnisse (Weiand, 2024). Das probabilistische Record-Linkage wurde mithilfe des Python-Pakets „recordlinkage“ durchgeführt (de Bruin et al., 2023).⁴ Als Schwellenwert für die Match-Wahrscheinlichkeit wurden – wie bei vorherigen Studien (Schnell, 2022) – 80 Prozent gewählt. Liegen für ein Record mehrere mögliche Paare vor, wurde nur das Paar mit der höchsten Wahrscheinlichkeit gewählt.

⁴ Beim verwendeten probabilistischen Record-Linkage-Modell handelt es sich um das im technischen Bericht als modifizierte Variante bezeichnete Modell (Weiand, 2024).

Im Gegensatz zum probabilistischen Verfahren erfolgt die Klassifikation beim deterministischen Verfahren durch die exakte Übereinstimmung einer Teilmenge von Merkmalen (Matchkey). Das implementierte Verfahren orientiert sich an ähnlich strukturierten Verfahren, die unter anderem beim Office of National Statistics verwendet werden (Bernstam et al., 2022; Shipsey & Plachta, 2021). Eine Klassifikation als Match erfolgt, sobald mindestens einer der gebildeten Matchkeys übereinstimmt.⁵ Liegen für ein Record mehrere mögliche Paare vor, erfolgt die Klassifikation anhand der Anzahl übereinstimmender Matchkeys.

Die Verknüpfung erfolgt jeweils zweistufig. In der ersten Stufe werden zunächst alle Fälle verknüpft, die auf der Menge der im Szenario gegebenen Personenmerkmale exakt übereinstimmen. Anschließend erfolgt die Verknüpfung der verbleibenden Fälle mit den genannten Record-Linkage-Methoden. Wie in der amtlichen Statistik üblich, wurden die Parameter für beide Methoden so gewählt, dass möglichst wenige falsch-positive Verknüpfungen entstehen (Schnell, 2019, S. 11). Dadurch wird verhindert, dass die Bildungsverläufe zweier unterschiedlicher Personen verknüpft werden.

3.4 Analyseverfahren

Die Qualität der Datenverknüpfung wurde anhand von Precision, Recall und dem F*-Maß (Hand et al., 2021) evaluiert. Precision misst den Anteil korrekter Verknüpfungen, Recall den Anteil aller gefundenen korrekten Verknüpfungen. Der Wert F* gewichtet beide Werte gleichermaßen.

Zusätzlich werden logistische Regressionen berechnet, um zu prüfen, ob bestimmte Gruppen von Schülerinnen und Schülern eine geringere Verknüpfungswahrscheinlichkeit aufweisen. Als abhängige Variable wurde der Verknüpfungserfolg genutzt. Unabhängige Variablen sind Geschlecht, Schulform, Herkunft, Bevölkerungsdichte und sozioökonomischer Status des Wohnorts (RISE-Index). Positive Koeffizienten zeigen eine Verbesserung und negative eine Verschlechterung des Verknüpfungserfolgs an. Aufgrund der unterschiedlichen Meldewege des ZSR und der damit einhergehenden Qualitätsunterschiede (Weiland, 2024) wurde kontrolliert, ob die Daten aus dem Melderegister oder von der Schule stammen.

Insbesondere die Einträge von Berufsschülerinnen und -schülern und von nicht in Hamburg lebenden Schülerinnen und Schülern stammen sehr häufig von den Schulen. Diese beiden Gruppen wurden daher von sämtlichen Analysen ausgeschlossen, da die Effekte beider Gruppen nicht ausreichend vom Effekt des Meldewegs getrennt werden können. Deskriptive Angaben über die in der Regression verwendeten Variablen finden sich in Tabelle 2.

5 In Szenario A wurden vier Matchkeys gebildet; in Szenario B zehn (Weiland, 2024, S. 64).

Tab. 2: Anzahl und Anteil Beobachtungen der unabhängigen Variablen im Schuljahr 2021/22

Variable	Ausprägungen	n	%
Geschlecht	Männlich	99.724	51,1
	Weiblich	95.245	48,9
Herkunft	Deutsch	130.717	67,0
	Migrationshintergrund	42.531	21,8
	Migrant	21.721	11,1
Datenquelle	Melderegister	186.918	95,9
	Schule	8.051	4,1
RISE-Index	Hoher Status (3–4)	153.535	78,7
	Niedriger Status (1–2)	41.434	21,3
Schultyp	Grundschule	68.125	34,9
	Stadtteilschule	64.617	33,1
	Gymnasium	56.609	29,0
	Andere	5.618	2,9

Quelle: eigene Berechnung.

4 Empirische Befunde

Tabelle 3 stellt die Linkage-Ergebnisse der beiden Szenarien A und B dar. Zusätzlich wird die Qualität der exakten Verknüpfung als Referenzgröße ausgewiesen. Die Qualität der exakten Verknüpfung zeigt den Anteil der Personen, die zu beiden Zeitpunkten exakt dieselben Merkmalsausprägungen aufweisen. Der Anteil der Personen mit Merkmalsveränderungen liegt bei etwa 11 Prozent. Die Ergebnisse verdeutlichen, dass beide Record-Linkage-Verfahren eine hohe Verknüpfungsqualität erzielen. Die Nutzung der Gitterzellen-Koordinate (Szenario B) führt zu einer leichten Verbesserung des Verknüpfungserfolgs. Insgesamt erzielt das probabilistische Record-Linkage im Szenario B die besten Verknüpfungsergebnisse.

Tab. 3: Linkage-Qualität der getesteten Methoden in den beiden Szenarien (SZ)

SZ	Methode	TP	FP	FN	Precision	Recall	F*
A	Exakt	162360	0	20025	1,000	0,890	0,890
	Prob. RL	181740	338	645	0,998	0,996	0,995
	Matchkeys	181683	60	702	1,000	0,996	0,996
B	Exakt	163340	0	19045	1,000	0,896	0,896
	Prob. RL	182005	92	380	0,999	0,998	0,997
	Matchkeys	181759	57	626	1,000	0,997	0,996

Quelle: eigene Berechnung.

Anm.: Die Methode Exakt beschreibt den ersten Verknüpfungsschritt, die exakte Verknüpfung anhand aller gegebener Merkmale (TP = true positive; FP = false positive; FN = false negative).

Die Ergebnisse der Regression werden in Tabelle 4 dargestellt. Dargestellt werden Average Marginal Effects (AMEs; Kohler & Kreuter, 2017). Sie lassen sich als durchschnittlicher Effekt einer Variable auf die Verknüpfungswahrscheinlichkeit interpretieren. Da diese Variablen lediglich einen indirekten Einfluss auf den Ausgang einer Verknüpfung haben, ist eine niedrige Modellgüte zu erwarten, denn nur die tatsächlichen Ausprägungen der Personenmerkmale wirken sich direkt aus.

Tab. 4: Einfluss soziodemografischer Merkmale auf den Erfolg einer Verknüpfung

Merkmal	Szenario A		Szenario B	
	Prob. RL	Matchkeys	Prob. RL	Matchkeys
Weiblich	0,034 (0,037)	0,003 (0,040)	0,009 (0,020)	0,031 (0,026)
Herkunft				
Deutsch				
Migrationshintergrund	-0,146*** (0,043)	-0,297*** (0,048)	-0,033* (0,019)	-0,166*** (0,028)
Migrant	-1,200*** (0,095)	-1,401*** (0,101)	-0,646*** (0,066)	-1,163*** (0,086)
Schulform				
Grundschule				
Stadtteilschule	-0,112** (0,046)	-0,057 (0,049)	-0,048** (0,023)	-0,062** (0,030)
Gymnasium	0,144*** (0,046)	0,226*** (0,049)	0 (0,027)	0,045 (0,034)
Andere	-0,084 (0,111)	0,008 (0,115)	-0,029 (0,056)	0,001 (0,068)
Wohnort (RISE)				
Hoher Status (3–4)				
Niedriger Status (1–2)	0,040 (0,044)	0,063 (0,046)	0,051** (0,022)	0,046 (0,028)
Anzahl Personen/Hektar	-0,004*** (0,001)	-0,005*** (0,001)	-0,002*** (0)	-0,002*** (0)
Schuldaten	-2,564*** (0,378)	-2,413*** (0,378)	-0,470*** (0,154)	-0,886*** (0,210)
Log Pseudo-Likelihood	-5.635,109	-6.160,288	-1.814,870	-2.770,701
χ^2	655,464	731,765	414,328	745,004
R^2 (Nagelkerke)	0,057	0,058	0,104	0,121
n	165.530	165.509	165.548	165.522

Quelle: eigene Berechnung.

Anm.: Dargestellt sind die AMEs in Prozent sowie deren Standardfehler in Klammern.

* $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.

Die Ergebnisse verdeutlichen einen migrations- und schulformbezogenen Linkage-Bias. Migrantinnen und Migranten können schlechter verknüpft werden als Deutsche. Der Besuch einer Stadtteilschule hat im Vergleich zu einer Grundschule einen negativen Effekt auf den Verknüpfungserfolg, während der Besuch eines Gymnasiums einen positiven Effekt hat. Die Bevölkerungsdichte hat ebenfalls einen negativen Einfluss auf den Verknüpfungserfolg, was bedeutet, dass Menschen, die in Großwohnsiedlungen oder Großgebäuden leben, schlechter verknüpft werden können. Dieser Effekt ist im Szenario A stärker als im Szenario B, was darauf hinweist, dass schwer zu verknüpfende Fälle vorzugsweise in Großwohnsiedlungen oder Großgebäuden wohnen. Das Geschlecht und der sozioökonomische Status des Wohnorts (RISE-Index) haben hingegen keinen eindeutigen Effekt auf den Verknüpfungserfolg.

Insgesamt zeigen sich in Szenario B schwächere Effekte als in Szenario A. Abgesehen von der Stärke des Effekts unterscheiden sich die Methoden und Szenarien jedoch nicht in der Richtung der Effekte. Die Methoden führen somit keinen zusätzlichen Linkage-Bias ein, der sich nicht über die unterschiedliche Linkage-Qualität erklären lässt. Auch auf Aggregatenebene der Stadtteile weisen die unterschiedlichen Methoden keinen zusätzlichen Linkage-Bias auf (Weiland, 2024, S. 74).

5 Diskussion der Ergebnisse

Den größten Erfolg bei der Verknüpfung von Daten aus den beiden Schuljahren erzielten wir auf Basis der Merkmale Vorname, Nachname, Geschlecht, Geburtsdatum und 100m-Gitterzellen-Koordinate. Ein besonders wichtiger Aspekt ist die Einbeziehung der 100-Meter-Gitterzelle, die die Notwendigkeit der exakten Anschrift überflüssig macht und somit einen höheren Datenschutz im Sinne der Datensparsamkeit bietet.

Die Ergebnisse zur Belastbarkeit der Datenverknüpfungen zeigen, dass insbesondere Schülerinnen und Schüler mit Migrationshintergrund eine erhöhte Wahrscheinlichkeit für fehlerhaft rekonstruierte Bildungsverläufe haben. Gleiches gilt bei Umzügen und Namensänderungen. Zudem variiert die Wahrscheinlichkeit für Verknüpfungsfehler regional. Da diese Befunde auf Daten aus einem kleinflächigen Stadtstaat basieren, ist anzunehmen, dass sich die damit einhergehenden Herausforderungen bei einer großflächigen Etablierung des vorgeschlagenen Vorgehens weiter potenzieren. Die Wahrscheinlichkeit für Verwechslungen anhand ähnlicher oder gleicher Personenmerkmale ist in einem bundesweiten BVR höher (Schnell, 2022). Daher bleibt der Geburtsort, zusammen mit den hier verwendeten Merkmalen, ein notwendiges Merkmal für die Erstellung einer Bildungs-ID.

Werden die Ergebnisse vor dem Hintergrund eines bundesweiten BVRs diskutiert, so ist zu beachten, dass die Verknüpfung eines kompletten (schulischen) Bildungsverlaufs nicht nur zwei, sondern weitaus mehr Jahre umfasst. Entsprechend bewirken bereits geringe Unterschiede in der Linkage-Qualität bei einer Verknüpfung (Tab. 3) einen erheblichen Unterschied, wenn komplette Bildungsverläufe abgebildet werden. Insbesondere systematische Fehler addieren sich dabei mit der Zeit.

Die Ergebnisse zeigen, dass selbst unter den besten Voraussetzungen keine vollständige Verknüpfung erreicht werden kann. Eine Vertrauensstelle – ähnlich wie in dem Krebsre-

gisterverfahren (Stegmaier et al., 2019) – wird daher Zweifelsfälle manuell prüfen müssen. Ferner können Fehlverknüpfungen auch über die gebildeten Bildungsverläufe identifiziert werden. Dabei können auffällige Brüche, unvollständige Lebensabschnitte oder fehlende Qualifikationen Aufschluss über eine Fehlverknüpfung bieten.

6 Zusammenfassung

Ziel der Studie war die Untersuchung der Rekonstruierbarkeit von Bildungsverläufen anhand von Personenmerkmalen. Wir analysierten, welche Merkmale für eine belastbare Datenverknüpfung notwendig sind und ob bestimmte Bevölkerungsgruppen dadurch verzerrte Bildungsverläufe aufweisen.

Wir konnten zeigen, dass mittels eines probabilistischen Record-Linkages auf Basis von Vorname, Nachname, Geschlecht, Geburtsdatum und 100m-Gitterzellen-Koordinate eine ausreichend hohe Verknüpfungsqualität erzielt werden kann. Aufgrund der Größenordnung eines bundesweiten BVRs ist jedoch davon auszugehen, dass ein zusätzliches, zeitlich stabiles Merkmal, wie der Geburtsort, für Datenverknüpfungen erforderlich ist (Schnell, 2019; Weiland, 2024).

Aus unseren Ergebnissen folgt, dass im Zuge der Etablierung eines BVRs – neben der Schaffung einer geeigneten Infrastruktur und gesetzlicher sowie datenschutzrechtlicher Regelungen – ein besonderer Fokus auf die Erhöhung der Datenqualität gelegt werden sollte. Dies betrifft insbesondere die für Record-Linkage zentralen Qualitätsdimensionen der Vollständigkeit und Genauigkeit der Angaben (Herzog et al., 2007). Das Risiko systematischer Ausfälle – sei es auch nur an bestimmten Schulformen oder in bestimmten Klassenformen – birgt ansonsten das Risiko systematischer Fehlschlüsse. Das ist besonders gravierend, wenn Kausalanalysen auf Basis der zu schaffenden Datengrundlage durchgeführt werden.

Unsere Analysen zeigen auch, dass es in amtlichen Daten Qualitätsunterschiede gibt. Die Qualität der Daten aus den Meldeämtern ist der Qualität der Daten aus den Schulen überlegen. Die Qualität letzterer scheint nicht ausreichend, um ein bundesweites BVR aufzubauen. Daraus folgt, dass eine Datenerfassung durch geschultes Personal zuverlässiger zu sein scheint. Nur wenn es gelingt, diese außerordentliche Datenqualität sicherzustellen und damit die Voraussetzungen für eine belastbare Grundlage für Datenverknüpfungen zu schaffen, wird es aus technischer Sicht möglich sein, ein BVR aufzubauen.

Neben diesen technischen Dimensionen ist die Etablierung eines BVR aber vor allem eine gestalterische Aufgabe. Sie hat tiefgreifende Implikationen für die Entwicklung von Schulen und Unterricht. Im Kontext aktueller bildungspolitischer Vorhaben wie dem Startchancen-Programm oder dem Digitalpakt Schule könnten mit Bildungsverlaufsdaten eine Grundlage geschaffen werden, um Mittel zielgenauer zu vergeben und Maßnahmen zu evaluieren. Mittels einer ebenenübergreifenden datengestützten Schul- und Unterrichtsentwicklung könnte ein BVR dadurch einen Beitrag leisten, die Effektivität des Bildungssystems maßgeblich zu steigern und Bildungsprozesse flexibler sowie individueller zu gestalten, vorausgesetzt, die technischen, ethischen und rechtlichen Rahmenbedingungen werden mit größter Sorgfalt definiert und umgesetzt.

Literatur und Internetquellen

- Behörde für Schule und Berufsbildung. (2022). *Hamburger Schulstatistik: Schuljahr 2021/22*. <https://doi.org/10.25654/IfBQ-BQ12:SCHULJAHRESERHEBUNG:2021-22>
- Behörde für Schule und Berufsbildung. (2024). *Hamburger Schulstatistik: Schuljahr 2023/24*. <https://doi.org/10.25654/IfBQ-BQ12:SCHULJAHRESERHEBUNG:2023-24>
- Behörde für Stadtentwicklung und Wohnen. (2022). *Rahmenprogramm Integrierte Stadtteilentwicklung: Leitfaden für die Praxis*. <https://www.hamburg.de/resource/blob/185456/7857e66072f972f6480c75e300de2698/leitfaden-rise-data.pdf>
- Bernstam, E. V., Applegate, R. J., Yu, A., Chaudhari, D., Liu, T., Coda, A., & Leshin, J. (2022). Real-World Matching Performance of Deidentified Record-Linking Tokens. *Applied Clinical Informatics*, 13 (4), 865–873. <https://doi.org/10.1055/a-1910-4154>
- Brändle, T. (2024, 2. Mai). Die Position: Eine Nummer fürs lebenslange Lernen. *DIE ZEIT*, 44. <https://www.zeit.de/2024/19/bildungs-id-digitalisierung-datenschutz-schueler>
- CDU, CSU, & SPD. (2025). *Verantwortung für Deutschland. Koalitionsvertrag zwischen CDU, CSU und SPD für die 21. Legislaturperiode des Deutschen Bundestages*. https://www.koalitionsvertrag2025.de/sites/www.koalitionsvertrag2025.de/files/koav_2025.pdf
- Christen, P., Ranbaduge, T., & Schnell, R. (2020). *Linking sensitive data: Methods and techniques for practical privacy-preserving information sharing*. Springer. <https://doi.org/10.1007/978-3-030-59706-1>
- de Bruin, J., Anderson, J., Becker, J., Wong, H., tylerbinski, Waleń, T., Elias, D., Weytjens, J., GG, D., Knuth, T., Varner, H., Baldassini, L., Hoffman, M., Antoine, M., Norton, R., Teigman, T., Deplasse, V., & andyessen. (2023). *J535D165/recordlinkage: V0.16 (Version v0.16)* [Software]. Zenodo. <https://doi.org/10.5281/zenodo.8169000>
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64 (328), 1183–1210. <https://doi.org/10.1080/01621459.1969.10501049>
- Fickermann, D. (2021). Daten für Taten. Verbesserung der Datengrundlagen für zielgerichteteres politisches Handeln zur Eindämmung und Bewältigung der Folgen der Corona-Pandemie. *DDS – Die Deutsche Schule*, 113 (2), 227–242. <https://doi.org/10.31244/dds.2021.02.09>
- Gawronski, K. (2020). Konzeption eines Bildungsregisters in Deutschland. *WISTA – Wirtschaft und Statistik*, 72 (2), 37–45. https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2020/02/konzeption-bildungsregister-022020.pdf?__blob=publicationFile
- Giar, K., Hohlstein, F., Wipke, M., & Scharnagl, A. (2023). Konzeption eines Statistischen Bildungsverlaufsregisters in Deutschland – Entwicklungen bis 2023 und Ausgestaltungsoptionen. *WISTA – Wirtschaft und Statistik*, 2023(3), 51–62.
- Hand, D. J., Christen, P., & Krielle, N. (2021). F*: An Interpretable Transformation of the F-Measure. *Machine Learning in Medicine*, 110 (3), 451–456. <https://doi.org/10.1007/s10994-021-05964-1>
- Hertweck, F., Isphording, I. E., Matthewes, S. H., Schneider, K., & Spieß, C. K. (2023). Bildungsdaten: Datenlücken durch ein Bildungsverlaufsregister schließen. *Wirtschaftsdienst*, 103(11), 733–736. <https://doi.org/10.2478/wd-2023-0204>
- Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer. <https://doi.org/10.1007/0-387-69505-2>
- Kohler, U., & Kreuter, F. (2017). *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung* (5. Aufl.). De Gruyter Oldenbourg. <https://doi.org/10.1515/9783110469509>
- Mundelius, M. (2019). Der Kerndatensatz auf der Basis von Individualdatenerhebungen in der Schulstatistik. Von Summendaten zu Einzeldaten. In D. Fickermann & H. Weishaupt (Hrsg.), *Bildungsforschung mit Daten der amtlichen Statistik* (DDS – Die Deutsche Schule, 14. Beiheft, S. 38–48). Waxmann. <https://doi.org/10.31244/dds.bh.2019.14.03>

- RatSWD (Rat für Sozial- und Wirtschaftsdaten). (2022). *Aufbau eines Bildungsverlaufsregisters: Datenschutzkonform und forschungsfreundlich* [RatSWD Positionspapier]. <https://www.kon-sortswd.de/wp-content/uploads/Positionspapier-RatSWD-Aufbau-eines-Bildungsverlaufsregisters.pdf>
- Schnell, R. (2019). *Eignung von Personenmerkmalen als Datengrundlage zur Verknüpfung von Registerinformationen im Integrierten Registerzensus* (No. WP-GRLC-2019-01; GRLC Working Paper Series). DuEPublico. <https://doi.org/10.17185/DUEPUBLICO/49551>
- Schnell, R. (2022). *Verknüpfung von Bildungsdaten in einem Bildungsregister mittels Record-Linkage auf Basis von Personenmerkmalen* (No. WP-GRLC-2022-03; GRLC Working Paper Series). DuEPublico. <https://doi.org/10.17185/duepublico/76331>
- Shipsey, R., & Plachta, J. (2021, 16. Juli). *Linking with anonymised data – how not to make a hash of it*. Office for National Statistics. <https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/linking-with-anonymised-data-how-not-to-make-a-hash-of-it>
- SWK (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz). (2022). *Entwicklung von Leitlinien für das Monitoring und die Evaluation von Förderprogrammen im Bildungsbereich* [Impulspapier]. https://www.swk-bildung.org/content/uploads/2024/02/SWK-2022-Impulspapier_Monitoring.pdf
- Statistische Ämter des Bundes und der Länder. (2004). Ergebnisse des Zensusstests. *WISTA – Wirtschaft und Statistik*, 2004 (8), 813–833.
- Stegmaier, C., Hentschel, S., Hofstädter, F., Katalinic, A., Tillack, A., & Klinkhammer-Schalke, M. (2019). *Das Manual der Krebsregistrierung*. W. Zuckschwerdt Verlag.
- Weiland, S. V. (2024). *Analyse der Fehler in Quasi-Identifikatoren in einem deutschen Schülerregister durch probabilistische Längsschnittverknüpfung* (No. WP-GRLC-2024-02; GRLC Working Paper Series). DuEPublico. <https://doi.org/10.17185/DUEPUBLICO/81929>

Tobias Brändle, Dr. habil., geb. 1984, Leiter der Abteilung Datenmanagement und -service sowie der Vertrauensstelle des Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ) Hamburg.

E-Mail: tobias.braendle@ifbq.hamburg.de

Korrespondenzadresse: IfBQ Hamburg, Beltgens Garten 25, 20537 Hamburg

ORCID: 0000-0001-8872-9872

Severin V. Weiland, M. A., geb. 1995, Wissenschaftlicher Mitarbeiter am Lehrstuhl für Methoden der Empirischen Sozialforschung an der Universität Duisburg-Essen.

E-Mail: severin.weiand@uni-due.de

ORCID: 000-0002-3203-469X

Rainer Schnell, Prof. em. Dr., geb. 1957, Professor für Methoden der Empirischen Sozialforschung an der Universität Duisburg-Essen.

E-Mail: sekretariat.schnell@uni-due.de

ORCID: 0000-0001-7843-4974

Korrespondenzadresse: Universität Duisburg-Essen, Lotharstr. 65, 47057 Duisburg