

Harmonisierung und Zusammenführen der Daten des Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) mit den Daten der Startkohorte 4 des Nationalen Bildungspanels (NEPS) – CILS4NEPS

Lena Arnold, Andreas Horr, Marlene Bauer,
Brigitte Finke, Jörg Dollmann

Juni 2026

Harmonisierung und Zusammenführen der Daten des Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) mit den Daten der Startkohorte 4 des Nationalen Bildungspanels (NEPS) – CILS4NEPS

Lena Arnold^{1*}, Andreas Horr², Marlene Bauer¹,
Brigitte Finke², Jörg Dollmann^{1, 3}

Juni 2026

<https://doi.org/10.5281/zenodo.1912908>

¹ Mannheimer Zentrum für Europäische Sozialforschung MZES

² Leibniz-Institut für Bildungsverläufe LIfBi

³ Deutsches Zentrum für Integrations- und Migrationsforschung DeZIM

Danksagung: Wir danken Christian Aßmann, Timo Gnambs, Anne Konrad, Kathrin Lockl, Verena Ortmanns, Ranjit Singh, Silke Schneider und Ariane Würbach für ihre wertvolle Beratung und Unterstützung im Rahmen der CILS4NEPS Datenharmonisierung.

Abstract

In dem vorliegenden Working Paper wird das Harmonisierungsprojekt CILS4NEPS vorgestellt, welches die Daten des Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) mit den Daten des National Bildungspanels Startkohorte 4 (NEPS SC4) verknüpft. Im Rahmen einer Ex-post Harmonisierung eröffnet CILS4NEPS durch die Kombination der beiden Datenquellen neue Forschungspotenziale in der Bildungs- und Integrationsforschung, die mit den jeweils einzelnen Datensätze nicht oder nur bedingt möglich wären. Das Working Paper dokumentiert systematisch den Prozess des Ex-post Harmonisierung; beschreibt und begründet dabei die eingesetzten Verfahren und illustriert diese anhand ausgewählter Beispiele; und skizziert die im Datenangebot vorgesehenen Begleitmaterialien hinsichtlich ihrer Funktion für Nachvollziehbarkeit und Replizierbarkeit.

Keywords: CILS4EU; NEPS SC4; Datenharmonisierung; Bildung; Migrant:innen; Paneldaten

Inhaltsverzeichnis

1. Einleitung	3
2. Datenbasis	4
2.1. CILS4EU.....	4
2.2. NEPS SC4	4
2.3. Vergleichbarkeit	5
3. Ergebnisbericht: Harmonisierung der Daten aus CILS4EU und NEPS SC4.....	5
3.1. Überblickserstellung	6
3.2. Kodierungsüberblick.....	8
3.3. Theoretische Erstellung des Zielitems	10
3.4. Datenaufbereitung CILS4EU und NEPS SC4	14
3.5. Erstellung der Zielitems in den Daten.....	14
3.6. Harmonisierung der Leistungsmessungen	18
3.7. Umgang mit Klassifikations-Heuristiken in CILS4EU und NEPS SC4.....	20
3.8. Erstellung des harmonisierten Datensatzes.....	20
3.9. Gewichtung	21
3.10. Erstellung der Dokumentation	21
3.11. Datenarchivierung und -zugang.....	22
4. Zusammenfassung und Fazit.....	22
Literaturverzeichnis	24

1. Einleitung

In modernen Gesellschaften zählen die Integration von Menschen mit Zuwanderungsgeschichte sowie die damit verbundenen Chancen und Herausforderungen zu den zentralen Fragen – wobei die Fragen nach den Bedingungsfaktoren des Bildungserfolgs eine Schlüsselrolle spielt. Um diese Fragen beantworten zu können, braucht es hochwertiges und detailliertes Datenmaterial. Mit dem Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU; Kalter et al., 2016) und dem Nationalen Bildungspanels (nachfolgend NEPS bzw. NEPS SC4, wenn auf die Startkohorte 4 Bezug genommen wird; Blossfeld, Rossbach & Maurice, 2011) stehen in Deutschland zwei zentrale Datensätze zur Verfügung, die diese Prozesse im Bildungsbereich in hoher Detailschärfe abbilden. Während CILS4EU durch seine internationale Ausrichtung und ein Oversampling von Schulen mit hohem Migrantenanteil besonders für die Analyse von Integrationsprozessen geeignet ist, bietet das NEPS eine detaillierte Erfassung von Bildungsverläufen und Übergängen in Ausbildung, Studium und Arbeitsmarkt.

In diesem Working Paper wird ein Harmonisierungsprojekt genauer beleuchtet, dessen Ziel es war, zusätzliche Forschungspotentiale durch die Kombination der beiden Datenquellen zu erschließen, die mit den jeweils einzelnen Datensätze nicht oder nur bedingt möglich wären (Dollmann et al., 2025). So stellt eine Kombination beider Quellen für national angelegte Analysen eine sinnvolle Bereicherung dar, da hierüber Fallzahlen sowohl für bestimmte Gruppen (ethnisch oder sozial) als auch für bestimmte Ereignisse (Übergang in bestimmte Schul- bzw. Ausbildungsformen) erhöht werden können, die in der Folge differenziertere Analysen ermöglichen als nur mit einem der beiden Datensätze alleine. Darüber hinaus ermöglicht eine Kombination der beiden Datensätze, dass die Daten von NEPS SC4 für internationale Vergleiche nutzbar gemacht werden können, um Schul- und Ausbildungskarrieren von Jugendlichen in Deutschland denen in England, den Niederlanden oder Schweden (den drei – neben Deutschland – teilnehmenden Ländern in CILS4EU) gegenüberzustellen. Neben Befragungsdaten der Schüler:innen und ihrer Eltern umfassen die harmonisierten Daten auch Kompetenzmessungen der Schüler:innen.

Das vorliegende Working Paper verfolgt drei Ziele. *Erstens* dokumentiert es den Prozess der Ex-post-Harmonisierung von CILS4EU und NEPS SC4 – von der Identifikation inhaltlich vergleichbarer Variablen bis zur Erstellung eines konsistenten, dokumentierten Datensatzes. *Zweitens* beschreibt und begründet es die eingesetzten Verfahren – von direkter Zuordnung und Rekodierung über Linear Equating bis hin zu Equipercentile Equating bei den Leistungsmessungen – und illustriert diese anhand ausgewählter Beispiele. *Drittens* skizziert es die vorgesehenen Begleitmaterialien (u. a. Codebuch, Kodierungs- und Überblickstabellen, Do-Files) im Datenangebot und erläutert deren Funktion für Nachvollziehbarkeit und Replizierbarkeit.

Der Beitrag des Working Papers liegt damit vor allem in Transparenz und Nachnutzbarkeit: Forschende können die getroffenen Entscheidungen prüfen, ihre Implikationen für eigene

Analysen einschätzen und die bei der Datenübermittlung bereitgestellten Materialien unmittelbar verwenden. Ziel ist es, die Nutzung bestehender Datenbestände zu erleichtern und die praktische Arbeit mit harmonisierten Umfragedaten zu unterstützen.

Im Folgenden wird zunächst die Datenbasis des Harmonisierungsprojektes vorgestellt (Kapitel 2). Anschließend werden das methodische Vorgehen und die zentralen Arbeitsschritte des Harmonisierungsprozesses erläutert (Kapitel 3), bevor mit einer Zusammenfassung und einem Fazit geschlossen wird (Kapitel 4).

2. Datenbasis

2.1. CILS4EU

CILS4EU ist eine internationale Längsschnittstudie zur Integration von Jugendlichen mit und ohne Migrationshintergrund in Deutschland, England, den Niederlanden und Schweden. Das Studiendesign orientiert sich an NEPS SC4; die Stichprobe wurde explizit stratifiziert mit Oversampling migrantenreicher Schulen und zusätzlich implizit – in Deutschland nach Bundesland und Schulart –, um diese Merkmale proportional zur Grundgesamtheit zu berücksichtigen.

In Welle 1 wurden rund 19 000 Jugendliche befragt (davon ca. 5 000 in Deutschland), etwa die Hälfte mit Migrationshintergrund. 2011–2013 folgten zwei weitere Wellen: Welle 2 erneut im Schulkontext, Welle 3 im Individualfeld (online/telefonisch). Nach Welle 3 wurde der deutsche Studienteil in das DFG-Langfristprogramm übernommen. Im Jahr 2016 (Welle 6) kam eine Auffrischungstichprobe hinzu, wiederum mit überproportionalem Anteil Jugendlicher mit Migrationshintergrund. Aktuell liegen neun Wellen und eine COVID-19-Sondererhebung vor; die zehnte Welle wird im Sommer 2026 veröffentlicht.

2.2. NEPS SC4

Das NEPS erhebt Längsschnittdaten zu Kompetenzentwicklungen, Bildungsprozessen, Bildungsentscheidungen, Bildungsrenditen in formalen, nicht-formalen und informellen Kontexten. Die Erhebungen 2008–2013 erfolgten im Rahmenprogramm zur Förderung der empirischen Bildungsforschung (BMBF); seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (IIfBi) in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt. Die Teilstudie SC4 setzt in Jahrgangsstufe 9 an und verfolgt Wege in und durch die Sekundarstufe II sowie Übergänge in berufliche Ausbildung, Studium, Arbeitsmarkt. Zielpopulation sind Schülerinnen und Schüler an Regel- und Förderschulen, die 2010/11 Klasse 9 besuchten; gezogen wurde eine geschichtete Klumpenstichprobe an Regelschulen sowie eine separate Förderschulstichprobe.

Zum Start waren die Jugendlichen 14–15 Jahre alt; in Welle 1 liegen für über 15 500 Personen Befragungsdaten vor, rund 37 % mit Migrationshintergrund. Jugendliche, die weiterhin die ausgewählten Schulen besuchten, wurden in den folgenden Wellen im Schulkontext befragt. Schulabgänger überwiegend per CATI im Individualfeld verfolgt. Die Haupterhebung an

Schulen erfolgte als PAPI durch IEA-DPC, CATI/CAWI im Individualfeld durch infas. Aktuell liegen für SC4 16 Wellen sowie eine COVID-19-Sondererhebung vor.

2.3. Vergleichbarkeit

Die Kombination der beiden Datensätze bietet sich an, da sich sowohl CILS4EU als auch NEPS SC4 auf dieselbe Zielpopulation (Jugendliche ab dem Alter von 14–15 Jahren) bzw. im Fall der deutschen Teilstudie von CILS4EU sogar auf dieselbe Grundgesamtheit beziehen (Jugendliche in Klasse 9 im Schuljahr 2010/2011) und CILS4EU einen sehr ähnlichen Samplingansatz auf Schulebene wie NEPS SC4 umgesetzt hat, dabei aber Schulen mit hohen Migrant:innenanteilen überproportional häufig gezogen wurden, während NEPS SC4 Schüler:innen, die eine Förderschule mit Schwerpunkt Lernen besuchen, bei der Erfassung kognitiver Grundfähigkeiten überproportional berücksichtigte (Nusser & Messingschlager, 2018). NEPS SC4 wurde in Deutschland bundesweit erhoben, CILS4EU bundesweit mit Ausnahme von Bayern. Das internationale Stichprobenkonzept von CILS4EU sowie die nationale Stichprobenziehung und Feldarbeit in den gezogenen Schulen in Deutschland wurde vom IEA-DPC (verantwortlich u.a. auch für PISA, TIMSS etc.) durchgeführt, das ebenfalls verantwortlich für das Stichprobenkonzept und große Teile der Feldarbeit bei NEPS SC4 war, was zusätzlich die Vergleichbarkeit und damit die sinnvolle Kombination der beiden Datenquellen gewährleistet. Wichtig ist, dass die Schulen, die für die NEPS SC4-Stichprobe ausgewählt wurden, direkt aus der deutschen CILS4EU-Stichprobe ausgeschlossen wurden. Dies bedeutet, dass dieselben Schüler im harmonisierten Datensatz nicht doppelt vorkommen.

3. Ergebnisbericht: Harmonisierung der Daten aus CILS4EU und NEPS SC4

Dieser Abschnitt dient dem Überblick über die jeweiligen Prozessschritte des Harmonisierungsverfahrens der Daten aus CILS4EU und NEPS SC4. Grundlage der Harmonisierung sind die Befragungen der Zielpersonen (Welle 1 bis 5 für CILS4EU und Welle 1 bis 8 für NEPS SC4, sowie Covid-19 Sonderbefragungen in beiden Erhebungen), sowie der Eltern. Da in CILS4EU die Elternbefragung nur in der ersten Welle durchgeführt wurde, nutzen wir zur bestmöglichen zeitlichen Passung auch in NEPS SC4 ausschließlich die Elternbefragung aus Welle 1. Lehrkräftebefragungen wurden im Rahmen dieses Vorhabens nicht berücksichtigt.

Im Gegensatz zur Ex-ante Harmonisierung von Daten, bei welcher Umfragen bereits vor ihrer Erhebung so konzipiert werden, dass sie vergleichbar sind, bezieht sich die Ex-post Harmonisierung auf die Harmonisierung bereits vorhandener Erhebungsdaten in einen integrierten Datensatz (Granda et al., 2010). Im Fall der CILS4NEPS- Harmonisierung ist die Ex-post Harmonisierung naheliegend, da in beiden Studien bereits über einen gewissen Zeitraum Daten erhoben wurden. Das Ziel der Ex-post Harmonisierung besteht darin, einen kombinierten Datensatz mit harmonisierten Variablen zu erstellen, die aus verschiedenen Quell-Datensätzen stammen, aber auf einer gemeinsamen Definition der Konstruktion

aufbauen (Wolf et al., 2016). Diese Kombination von Datensätzen kann sowohl für länderübergreifende als auch für nationale Erhebungen durchgeführt werden.

Es existieren keine fest etablierten Schritte für die Ex-post Harmonisierung von Daten, jedoch schlagen Experten folgenden Schritte vor (Granda et al., 2010; Singh, 2021):

- 1) Identifizierung der zu kombinierenden Datensätze
- 2) Identifizierung ähnlicher Fragen in den Quell-Fragebögen, die Potenzial für eine Harmonisierung bieten
- 3) Definition von Zielitems, die die Quell-Variablen zu harmonisierten Variablen kombinieren
- 4) Definition und Entscheidung über Harmonisierungsstrategien zur Erstellung der Zielitems
- 5) Abbildung der während der Datenharmonisierung angewandten Routinen zur Gewährleistung der Replizierbarkeit.

Wir haben uns an diesen Schritten für die Harmonisierung von CILS4EU und NEPS SC4 orientiert und erläutern im Folgenden, wie jeder dieser Schritte im Harmonisierungsprozess umgesetzt wurde. Zunächst wird dargestellt, wie die Befragungsdaten und die Leistungsmessungen (Abschnitte 3.1 bis 3.8) jeweils inhaltlich auf Überschneidungen analysiert und darauf basierend harmonisiert wurden. Abschnitte 3.9 bis 3.11 enthalten Informationen bezüglich Gewichtung, Dokumentation und Datenzugang.

3.1. Überblickserstellung

Im ersten Schritt der Harmonisierung von CILS4EU und NEPS SC4 wurde eine umfassende Überblickstabelle erstellt, die sämtliche relevanten Variablen aus den Schüler:innenbefragungen (CILS4EU Wellen 1–5; NEPS SC4 Wellen 1–8; einschließlich der COVID-19-Sondererhebungen) sowie aus den Elternbefragungen systematisch zusammenführt. Für jede erfasste Variable sind Variablenname, Frageformulierung, Antwortkategorien, die jeweils einbezogenen Befragtengruppen und die Wellen dokumentiert. Abbildung 1 zeigt einen exemplarischen Ausschnitt der Tabelle.

CILS4EU (w1-w3)

Content Area	Indicator	Variable name	Question	Categories	Students Wave 1	Parents Wave 1	Students Wave 2	Students Wave 3
Socio-Demography and Migration								
General information								
General information		Sex			X	X	X	X
General information	Sex	sex	Students wave 1: Are you a boy or a girl? Parents wave 1: Are you male or female?	Boy (Parents wave 1: Male) Girl (Parents wave 1: Female)	X	X	X	X

NEPS Starting Cohort 4 = 9th Grade (w1-w10)

Variable name	Question	Categories													
			Wave 1		Wave 2		Wave 3		Wave 4	Wave 5				Wave 6	
			Regular schools	Special schools	Regular schools	Special schools	Graduates	Pupils	Graduates	Pupils		School leavers		School leavers	
			First-time respondents	First-time respondents	First-time respondents	First-time respondents	First-time respondents	First-time respondents	Panel respondents	First-time respondents	Panel respondents	First-time respondents	Panel respondent	Panel respondents	Pupils
Socio-Demography and Migration															
General information															
Sex															
t700031	Are you...	...male? ...female?	X	X	X	X	X (t700001)	X	X (t700001)	X	X	X (t00001)	X (t700001)	X (t700001)	X

Abbildung 1: Auszug aus der Überblickstabelle für die Variable „Geschlecht“ in CILS4EU und NEPS SC4

Die Zuordnung ist so angelegt, dass auf der rechten Tabellenseite alle CILS4EU-Variablen stehen, während die linke Seite die inhaltlich potenziell entsprechenden Variablen aus NEPS SC4 ausweist. Dadurch wird auf einen Blick erkennbar, wo inhaltliche Übereinstimmungen bestehen, wo ähnliche Konstrukte erfasst werden und wo keine Entsprechungen vorliegen. Um diese Äquivalenzen transparent zu klassifizieren, wird ein Farbschema verwendet: Grün markierte Paare sind (nahezu) identisch hinsichtlich Fragestellung und Antwortkategorien, Gelb markierte Paare erfassen ähnliche, aber nicht vollständig deckungsgleiche Konstrukte, Rot kennzeichnet Variablen, die ausschließlich in einem der beiden Datensätze vorkommen.

Die Erstellung der Überblickstabelle erfolgte schrittweise. Zunächst wurden alle in CILS4EU (Wellen 1–5 sowie COVID-19-Sondererhebung) enthaltenen Variablen aufgelistet und gemäß dem CILS4EU-Codebuch in Inhaltsbereiche gegliedert, die später jeweils separat harmonisiert werden können. Anschließend wurden mit Hilfe der NEPS SC4-SUF-Fragebögen und des NEPS-Codebuchs inhaltliche Äquivalente gesucht und – sofern vorhanden – mit Variablennamen und Antwortkategorien eingetragen. Die Wellenabdeckung in NEPS SC4 wurde dabei wie folgt gekennzeichnet: „X“ wenn eine CILS4EU-äquivalente Frage gestellt wurde, „X(...)“ wenn die Frage gestellt wurde, der Variablenname in NEPS SC4 jedoch zwischen den Wellen abwich, sowie ein „.“ wenn keine annähernd vergleichbare Variable vorhanden war. Diese Markierungen erlauben eine schnelle Navigation über Wellen und Befragtengruppen hinweg.

Auf Basis dieser Klassifikation lässt sich das Harmonisierungspotenzial unmittelbar quantifizieren, indem die Zahl der grünen und gelben Variablen ermittelt wird. Insgesamt kommen damit in den Schüler:innenbefragungen rund 130 Variablen und in den Elternbefragungen 50 Variablen für eine Harmonisierung grundsätzlich in Frage. Die Überblickstabelle bildet damit nicht nur die Grundlage für die nachfolgenden Arbeitsschritte – insbesondere für die Ableitung und Definition der sogenannten Zielitems (d. h. der im harmonisierten Datensatz entstehenden Variablen aus der Kombination der CILS4EU- und NEPS SC4-Informationen) –, sondern erfüllt zugleich eine wichtige Dokumentationsfunktion. Sie stellt ein vollständiges Verzeichnis aller für die Harmonisierung relevanten Variablen dar und erhöht Transparenz und Nachvollziehbarkeit der getroffenen Entscheidungen.

3.2. Kodierungsüberblick

Zweck der Kodierungsübersicht in Form einer Excel-Tabelle ist es, eine präzise Übersicht darüber zu erhalten, wie die zu harmonisierenden Variablen aus CILS4EU und NEPS SC4 kodiert sind, um diese Kodierung anschließend bei der Harmonisierung der Variablen im Softwarepaket Stata (StataCorp., 2021) gezielt berücksichtigen zu können. Die Kodierungstabelle ist in nach Inhaltsbereichen gegliederte Arbeitsblätter aufgebaut, z. B. „Allgemeine Informationen“, „Haushaltssituation“. Sie bildet die operative Grundlage der Harmonisierung, da die Arbeit direkt auf ihr erfolgen kann, ohne fortwährend auf Codebücher oder Fragebögen zurückgreifen zu müssen. Zugleich dokumentiert sie die getroffenen Entscheidungen transparent.

Ausgehend von der Überblickstabelle werden die für die Harmonisierung relevanten Variablen aus CILS4EU und NEPS SC4 in die jeweils passenden, inhaltsbezogenen Excel-Blätter der Kodierungstabelle übernommen. Gleiche inhaltliche Informationen und gleiche Kodierungen werden in derselben Zeile geführt; Abbildung 2 veranschaulicht die Eintragung. Für die CILS4EU- und NEPS-SC4-Variablen ist festgehalten, welche Antwortkategorien wie kodiert werden, einschließlich der „Missings“. Bei NEPS SC4 ist zusätzlich vermerkt, in welchem Datensatz die Variable abgelegt ist, so dass eindeutig ist, welche NEPS-SC4-Datei später heranzuziehen ist. Zentral ist, dass identische Antwortkategorien – unabhängig von ihrer ursprünglichen Kodierung – in derselben Zeile stehen, so dass Übereinstimmungen der Antwortkategorien und etwaige Abweichungen in der Kodierung unmittelbar sichtbar werden; im Fall der Geschlechtsvariable sind die Kodierungen in CILS4EU und NEPS SC4 identisch, eine Rekodierung ist hier nicht erforderlich.

CILS4EU W2		NEPS SC4			
		"Standard Item"	W1		
			Regular schools	Special schools	
			First-time respondents	First-time respondents	
			Code Questionnaire	Code Questionnaire	
Item: Sex (Indicator)		Item: Sex (Indicator)			
Question number	1	Question number	241 : 1		
Variable	y1_sex Sex	Variable	t700031 Sex	t700031	t700031
Data record		Data record	pTarget		
Filter		Filter			
Filter_contentual		Filter_contentual			
Introduction		Introduction			
Question text	Are you a boy or a girl?	Question text	Are you...		
Submit list		Submit list			
Instruction		Instruction	«Please check where applicable.»		
Interviewer Instruction		Interviewer Instruction			
Answer categories:		Answer categories:			
Male	1 Male	Male	1 Male		
Female	2 Female	Female	2 Female		
Missing 1		Missing 1	-54 Missing by design		
Missing 2	-55 Other missing	Missing 2			
Missing 3		Missing 3			
Missing 4	-88 No answer	Missing 4			
Missing 5		Missing 5	-90 Unspecific missing		
Missing 6		Missing 6	-95 Implausible value		
Generated variable		Generated variable			
Number of response	2	Number of response	2		
Data remark		Data remark			

Abbildung 1: Auszug aus der Kodierungstabelle

Bei den Klassifizierungen der „Missings“ in CILS4EU und NEPS SC4 werden diese zunächst getrennt aufgeführt und am Ende in ein einheitliches Missing-Schema überführt. Ergänzend werden relevante Filterfragen in der Kodierungstabelle dokumentiert, da auch sie gegebenenfalls zu harmonisieren sind. Variablen, die in der Überblickstabelle rot markiert

wurden – also nur in CILS4EU vorhanden sind –, werden nicht in die Kodierungstabelle aufgenommen.

3.3. Theoretische Erstellung des Zielitems

Ziel dieses Arbeitsschrittes ist es, eine exakte Arbeitsgrundlage für die Erstellung der harmonisierten Zielitems zu schaffen und deren Zusammensetzung in der Kodierungstabelle transparent zu dokumentieren. Als Zielitems werden jene Variablen im harmonisierten Datensatz bezeichnet, die aus der Kombination der entsprechenden CILS4EU- und NEPS-SC4-Variablen entstehen. Anhand der in der Kodierungstabelle eingetragenen Zielitems lässt sich nachvollziehen, wie die jeweiligen Variablen rekodiert und zusammengeführt werden müssen – die zentrale Ausgangsbasis für die Erstellung der Zielitems in Stata.

Für jedes Zielitem werden zunächst eine Fragenummer und ein Variablenname vergeben, anschließend wird dokumentiert, aus welchen CILS4EU- und NEPS-SC4-Variablen es gebildet wird. Danach erfolgt eine Bewertung der Vergleichbarkeit dieser Variablen – sowohl inhaltlich im Hinblick auf das erfasste Konstrukt als auch in Bezug auf die Antwortkategorien. Grundlage dieser Beurteilung bildet einschlägige Forschungsliteratur zur Ex-post-Harmonisierung von Umfragedaten (z.B. Schneider, 2016; Wolf et al., 2016; Granda et al., 2010; Hoffmeyer-Zlotnik, 2008).

Im Zuge dieser Bewertung wird geprüft, ob die betreffenden Variablen aus CILS4EU und NEPS SC4 dasselbe Konstrukt erfassen und ob es sich dabei um ein beobachtbares (z. B. Bildungsstand, Haushaltsgröße) oder ein latentes, nicht direkt beobachtbares Konstrukt handelt. Anschließend werden die Antwortkategorien auf ihre inhaltliche und formale Ähnlichkeit sowie auf die Anzahl der Kategorien hin verglichen. Diese Analyse bildet die Grundlage für die spätere Auswahl der jeweils passenden Harmonisierungsstrategie in den Daten. Die Vergleichbarkeit der Variablen wird in der Kodierungstabelle mithilfe eines Farbschemas dokumentiert, das auch im harmonisierten Datensatz und im begleitenden Codebuch wiedergegeben wird. Während die Kodierungsübersicht sämtliche klassifizierte Variablen enthält, umfasst der CILS4NEPS-Datensatz lediglich jene Variablen, die als unproblematisch oder als komplexer, aber prinzipiell harmonisierbar eingestuft wurden.

Unproblematisch-klassifizierte Zielitems basieren auf Variablen, die in beiden Datensätzen als beobachtbare Konstrukte erfasst werden. Ein Beispiel ist der „Geburtsmonat“, der sowohl in CILS4EU als auch in NEPS SC4 mit der Frage „Wann sind Sie geboren, Monat?“ erhoben und von 1 („Januar“) bis 12 („Dezember“) kodiert wird (vgl. Abbildung 3). Aufgrund der Übereinstimmung in Konstrukt-Ähnlichkeit und Antwortkategorien kann hier ein einfaches Harmonisierungsverfahren in Form einer direkten Zuordnung beziehungsweise gegebenenfalls Re-Kodierung angewendet werden.

Item: Date of birth, Month		CILS4EU	NEPS SC4
2	Question number	2	241 : 2
H_dobm Date of birth, Month	Variable	y3_dobm Date of Birth, Month	t70004m Date of Birth, Month
CILS: y*_dobm	Data record		pTarget
NEPS: t70004m & t70000m			
	Filter		
	Filter_contentual		
	Introduction		
When were you born? Month	Question text	When were you born? Month	When were you born? < Month
	Submit list		
	Instruction		«Please enter the month and year as numbers and right-aligned.»
	Interviewer Instruction		
	Answer categories:		
1 January	January	1 January	1 January
2 February	February	2 February	2 February
3 March	March	3 March	3 March
4 April	April	4 April	4 April
5 May	May	5 May	5 May
6 June	June	6 June	6 June
7 July	July	7 July	7 July
8 August	August	8 August	8 August
9 September	September	9 September	9 September
10 October	October	10 October	10 October
11 November	November	11 November	11 November
12 December	December	12 December	12 December
-44 Interrupted interview	Missing 1	-44 Interrupted Interview	
-52 Implausible value removed	Missing 2		-52 Implausible value removed
-54 Missing by design	Missing 3		-54 Missing by design
-55 Other missing	Missing 4	-55 Other missing	
-66 Question not asked	Missing 5	-66 Question not asked	
-88 No answer	Missing 6	-88 No answer	
-90 Unspecific missing	Missing 7		-90 Unspecific missing
-95 Implausible value	Missing 8		-95 Implausible value
-99 Don't know	Missing 9		-98 Don't know
	Generated variable		
12	Number of response categories	12	12

Abbildung 2: Beispiel eines als *unproblematisch*-klassifizierten Zielitems – „Geburtsmonat“

Komplizierter-klassifizierte Zielitems beziehen sich zumeist auf latente Konstrukte, die in beiden Datensätzen inhaltlich ähnlich, aber nicht vollständig identisch gemessen werden. Mitunter handelt es sich auch um manifeste Konstrukte, deren Harmonisierung eine komplexere Zuordnung der Antwortkategorien erfordert (z. B. durch „lagging“ der Kategorien). Ein Beispiel ist das Konstrukt „Selbstwirksamkeit in der Schule“ (vgl. Abbildung 4), das in CILS4EU und NEPS SC4 auf leicht unterschiedliche Weise abgefragt wird – einmal auf einer fünf- und einmal auf einer vierstufigen Skala. Eine einfache Zuordnung der Antwortkategorien würde hier zu Verzerrungen führen. Diese Zielitems können von Datennutzer:innen grundsätzlich verwendet werden, sollten aber im Rahmen eigener Analysen validiert werden, etwa durch Überprüfung theoretisch erwartbarer Korrelationen (Singh, 2020, 2021; Kolen & Brennan, 2014). Für diese Variablen wird ein Harmonisierungsverfahren nach dem Prinzip des Linear Equating angewandt, das in Abschnitt 3.5 näher erläutert wird.

Item: Self-efficacy - do well at school	Item: Self-efficacy I	CILS4EU		NEPS	
36	Question number		16		740 : 55214i
H_sesch	Variable	y2_seff1	Self-efficacy I	t66002c	Self-concept school: good in most school subjects
CILS: y*_seff1	Data record				pTarget
	Filter				
	Filter_contentual				
	Introduction				
	Question text		How much do you agree or disagree with each of these statements? I am sure that I can do well at school.		[NCS] How are you doing in school? I'm good in most subjects.
	Submit list				
	Instruction				
	Interviewer Instruction				
	Answer categories:				
Rational .					
Rational Strongly agree	Strongly agree	1	Strongly agree	4	does completely apply
Rational .	Agree	2	Agree	3	does rather apply
Rational .	Neither agree nor disagree	3	Neither agree nor disagree		
numbers					
Rational .	Disagree	4	Disagree	2	does rather not apply
Rational Strongly disagree	Strongly disagree	5	Strongly disagree	1	does not apply at all
-44 Interrupted interview	Missing 1	-44	Interrupted interview		
-54 Missing by design	Missing 2			-54	Missing by design
-55 Other missing	Missing 3	-55	Other missing		
-66 Question not asked	Missing 4	-66	Question not asked		
-88 No answer	Missing 5	-88	No answer		
-90 Unspecific missing	Missing 6			-90	Unspecific missing
-95 Implausible value	Missing 7			-95	Implausible value
	Generated variable				
~	Number of response	5		4	

Abbildung 3: Beispiel eines als *komplizierter*-klassifizierten Zielitems – „Selbstwirksamkeit in der Schule“

Semi-problematisch-klassifizierte Zielitems messen in beiden Datensätzen latente Konstrukte, sind jedoch hinsichtlich der gemessenen Konstrukte und Antwortkategorien nur bedingt vergleichbar. Wie in dem untenstehenden Beispiel „Probleme mit Lehrern“ (vgl. Abbildung 5) verdeutlicht wird, messen die Fragen zwar ähnliche Konstrukte, jedoch werden in CILS4EU die Antwortkategorien in Form von Häufigkeiten und in NEPS SC4 in Form von Zustimmung gemessen. Diese Zielitems werden als „semi-problematisch“ klassifiziert. Sie werden in dieser Version von CILS4NEPS nicht harmonisiert.

Problematisch-klassifizierte Zielitems werden zwar dokumentiert, aber ebenfalls nicht harmonisiert. Zwar existieren in beiden Datensätzen Variablen, die ein ähnliches Konstrukt erfassen könnten, doch sind die Abweichungen in Fragestellung und Antwortformat so gravierend, dass keine ausreichende Konstruktvergleichbarkeit mehr gegeben ist. Ein Beispiel hierfür ist das Item „Geschlechterrollen“ (vgl. Abbildung 6), das in CILS4EU als *single-barrel* item mit Antwortoptionen nach Mann/Frau/Beiden und in NEPS SC4 als doppelt formulierte Zustimmungsfrage vorliegt. Aufgrund dieser Unterschiede wären deutliche Verzerrungen im harmonisierten Zielitem und in darauf basierenden Analysen zu erwarten, weshalb dieses Item nicht in den Datensatz aufgenommen wird.

CILS4EU					NEPS			
Deviance and Delinquency	Problem behavior in school I (wave 1); School-module: Problem behavior in school I (wave 3)	pbsch1 (students wave 3: s_pbsch1)	How often do you argue with teacher?	Every day Once or several times a week Once or several times a month Less often Never	t321813	And to what extent does the following statement apply? I often have problems or conflicts with my teachers.	does not apply at all does rather not apply does partly apply does rather apply does completely apply	

Abbildung 4: Beispiel eines als semi-problematisch-klassifizierten Zielitems (das Zielitem wurde nicht in die Kodierungstabelle eingetragen)

CILS4EU					NEPS			
Attitudes and Norms	Gender roles: Child care	gro1	In a family, who should do the following? Take care of the children	Mostly the man Mostly the woman Both about the same	t44613a	It's the man's job to earn money and the woman's job to take care of the household and family.	completely disagree rather disagree rather agree completely agree	max. ja/nein Frage erstellen? + Hinweis Probleme bei NEPS Frage

Abbildung 5: Beispiel eines als problematisch-klassifizierten Zielitems (das Zielitem wurde nicht in die Kodierungstabelle eingetragen)

3.4. Datenaufbereitung CILS4EU und NEPS SC4

Ziel dieses Arbeitsschrittes ist es, die Datensätze von CILS4EU und NEPS SC4 in ein einheitliches Format zu überführen, um anschließend die Harmonisierung im zusammengeführten Datensatz vornehmen zu können. Die Schüler:innenbefragungen von CILS4EU liegen ursprünglich im „weiten“ Format vor, während NEPS SC4 ein „langes“ Format nutzt. Im „weiten“ Format hat jede Person genau eine Datenzeile, in der die Informationen aus allen Wellen spaltenweise abgelegt sind; im „langen“ Format hat jede Person so viele Zeilen wie Wellen, da jede Welle eine eigene Zeile bildet. Zur Angleichung wurden die harmonisierten CILS4EU-Wellen kombiniert und in ein langes Format konvertiert. Da im weiten Format jede Variable ein Wellenpräfix trägt (z. B. „y1_“), kam es bei der Umstellung vereinzelt zu Anpassungen in der Beschriftung von Antwortkategorien, wenn sich über die Wellen hinweg kleine Abweichungen ergaben. Diese Änderungen an den Originaldaten sind im technischen Anhang zu CILS4NEPS dokumentiert (Dollmann et al., 2023). Zur Kennzeichnung der Herkunft erhielten alle CILS4EU-Variablen das Präfix „CILS4EU_“.

Für NEPS SC4 waren keine strukturellen Änderungen erforderlich. Die für die Harmonisierung relevanten Datensätze – „Cohort Profile“, „pTarget“, „pTargetCATI“, „Weights“ – wurden gemäß der vom NEPS-Datenzentrum bereitgestellten Merging-Matrix zusammengeführt; analog zu CILS4EU erhielten alle NEPS-Variablen das Präfix „NEPS_“, bevor beide Quellen zusammengeführt wurden. Bei den Elternbefragungen sowie der COVID-19-Zusatzbefragung war eine Formatumstellung wie bei den Schüler:innenbefragungen nicht notwendig.

3.5. Erstellung der Zielitems in den Daten

Nachdem die Zielitems theoretisch in der Kodierungstabelle definiert wurden, beschreibt dieser Abschnitt deren praktische Umsetzung in den Daten. Grundlage ist die zuvor eingeführte Klassifizierung in *unproblematische* und *kompliziertere* Zielitems: Erstere werden durch eine *direkte Zuordnung der Antwortkategorien* harmonisiert, letztere mithilfe eines Verfahrens des *Linear Equating*. Beide Vorgehensweisen werden im Folgenden kurz erläutert.

Die *Zuordnung der Antwortkategorien* zwischen den Quell-Variablen erforderte teils eine umgekehrte Kodierung der Antwortskalen oder die Zusammenfassung mehrerer Kategorien zu einer gemeinsamen Kategorie – abhängig von der jeweiligen Kodierung der Quell-Variablen. In anderen Fällen wurden mehrere Variablen aus CILS4EU oder NEPS SC4 zu einer harmonisierten Variable zusammengeführt. Ein Beispiel ist das Zielitem „Haushaltsmitglieder – Mutter, Stiefmutter, Adoptivmutter“: In NEPS SC4 werden verschiedene potenzielle Haushaltsmitglieder (z. B. Mutter, Stiefmutter, Adoptivmutter) in einer Frage erhoben, während CILS4EU diese Informationen in separaten Fragen erfasst. Entsprechend basiert das harmonisierte Zielitem auf einer NEPS-SC4-Variable, aber auf drei CILS4EU-Variablen.

Beim *Linearen Equating* wird davon ausgegangen, dass beide Werteverteilungen der zu harmonisierenden Instrumente näherungsweise normalverteilt sind. Dies impliziert, dass sich die beiden Verteilungen der Instrumentenwerte nur im Mittelwert und der Standardabweichung unterscheiden (Singh, 2020, 2021). Wichtig für die Anwendung dieses

Verfahrens ist, dass die zu harmonisierenden Instrumente aus derselben Grundgesamtheit gezogen wurden (Kolen & Brennan, 2014; Singh, 2020). Bei *Linear Equating* werden die Werte des Ausgangsinstruments linear transformiert, sodass der transformierte Mittelwert und die Standardabweichung des Ausgangsinstruments gleich dem Mittelwert und der Standardabweichung des Zielinstruments werden (Kolen & Brennan, 2014; Singh, 2020/2021). Durch dieses Verfahren haben die Befragten sehr ähnliche Punktwerte auf dem transformierten Ausgangsinstrument und dem Zielitem, je nach ihrer Position in der Normalverteilung. Befragte mit demselben z-Score haben denselben harmonisierten Wert, der allerdings auf das Skalen-Format des Zielitems skaliert ist (Singh, 2021: 128).

In der Anwendung des Linearen Equatings in der Datenharmonisierung wählten wir die CILS4EU-Variablen als Zielitems aus und setzten die Skala jedes NEPS SC4 Quellitems mit dem jeweiligen CILS4EU Zielitem gleich. Die lineare Transformation wurde in einem Excel-Spreadsheet durchgeführt, welches zu Zwecken der Dokumentation für jedes linear gleichgesetzte (equated) Zielitem zur Verfügung steht. Die Transformation resultiert in einer Tabelle, welche angibt, wie jede Antwortkategorie des jeweiligen NEPS SC4 Items rekodiert werden muss, um dem Zielitem zu entsprechen. Bei der Eintragung der Häufigkeitsverteilungen im Excel Spreadsheet beziehen wir ausschließlich die deutsche Befragtengruppe der CILS4EU-Erhebung mit ein. Der Grund hierfür ist, dass Lineares Equaten diese Zielpopulation erfordert, um eine zuverlässige Rekodierungstabelle zu erstellen. Diese kann jedoch anschließend auf die gesamte Stichprobe (auch nicht-deutsche Befragtengruppen in CILS4EU) angewendet werden. Das bedeutet, dass wir nicht nur die deutschen CILS4EU-Daten mit den NEPS SC4-Daten gleichsetzen, sondern auch die deutschen CILS4EU-Daten mit den CILS4EU-Daten aus England, den Niederlanden und Schweden. Wir gewichten die Daten während des linearen Gleichsetzungsprozesses zusätzlich mit den anwendbaren Individual-Gewichten der Datensätze (houwgt für CILS4EU; w_t für NEPS SC4), um eine valide Repräsentation der Stichprobenpopulationen zu erreichen (diese Gewichtung ist nur relevant, um die korrekte Häufigkeitsverteilung jedes Quell-Items zu erhalten, ist aber nicht im harmonisierten Item selbst enthalten). Wir gleichen auch Items an, die in der CILS4EU- und NEPS SC4-Erhebung in unterschiedlichen Jahren gemessen werden, geben diese Abweichungen jedoch in einer Flaggenvariablen an. Abbildung 7 zeigt die Eingabetabelle der Excel-Tabelle, in die die jeweiligen gewichteten Häufigkeiten für die Ziel- (CILS4EU) und Ausgangspositionen (NEPS SC4) eingetragen werden.

	A	B	C	I	E	F	G	H
1		CILS			NEPS			Variable name
2		Value	Count		Count	Value		H_spm
3		0	0		0	0		
4		1	598.67		3229.62	1		
5		2	1542.16		6017.36	2		
6		3	1975.42		2339.04	3		
7		4	728.23		1512.71	4		
8		5	159.62		0	5		
9		6	0		0	6		
10		7	0		0	7		
11		8	0		0	8		
12		9	0		0	9		
13		10	0		0	10		

Abbildung 6: Eingabetabelle

Auf Grundlage dieser Eingabetabelle werden die NEPS SC4-Antwortkategorien auf der rechten Seite linear transformiert, um sie an die CILS4EU-Antwortkategorien auf der linken Seite anzugleichen. Abbildung 8 zeigt die Rekodierungstabelle (siehe grüne Zahlen in der Mitte), die sich aus dieser linearen Transformation ergibt.

	A	B	C	F	I	J
1		H_spm				
2		CILS			NEPS	
3		Value	Count	MATCHED	Count	Value
5		0	0	0.00	0	0
6		1	598.67	1.44	3229.62	1
7		2	1542.16	2.49	6017.36	2
8		3	1975.42	3.54	2339.04	3
9		4	728.23	4.59	1512.71	4
10		5	159.62	0.00	0	5
11		6	0	0.00	0	6
12		7	0	0.00	0	7
13		8	0	0.00	0	8
14		9	0	0.00	0	9
15		10	0	0.00	0	10
17		Total	5004.1		13098.73	Total

Abbildung 7: Rekodierungstabelle

In diesem Beispiel zeigt die Rekodierungstabelle, dass die Antwortkategorie „1“ in den NEPS SC4-Daten für das harmonisierte Zielitem auf „1,44“ rekodiert werden muss, um der Antwortskala des CILS4EU-Items zu entsprechen. Abbildung 9 veranschaulicht das Rekodierungsverfahren grafisch. Die x-Achse stellt die Antwortkategorien des NEPS SC4-Datensatzes dar, während die y-Achse die CILS4EU-Antwortkategorien darstellt. Die gelbe Linie

zeigt die Mittelwerte der einzelnen Items (CILS4EU und NEPS SC4). Die grüne Linie stellt genau eine Standardabweichung vom Mittelwert der einzelnen Items dar. Die Rekodierungslinie (in blau) entsteht durch das Ziehen einer Linie durch die Schnittpunkte der gelben und grünen Linie. Um die in der obigen Rekodierungstabelle angegebenen Zahlenwerte zu erhalten, folgen wir den Zahlen der x-Achse („1“) nach oben bis zur blauen Linie und lesen den Rekodierungswert auf der y-Achse links ab („1,44“).

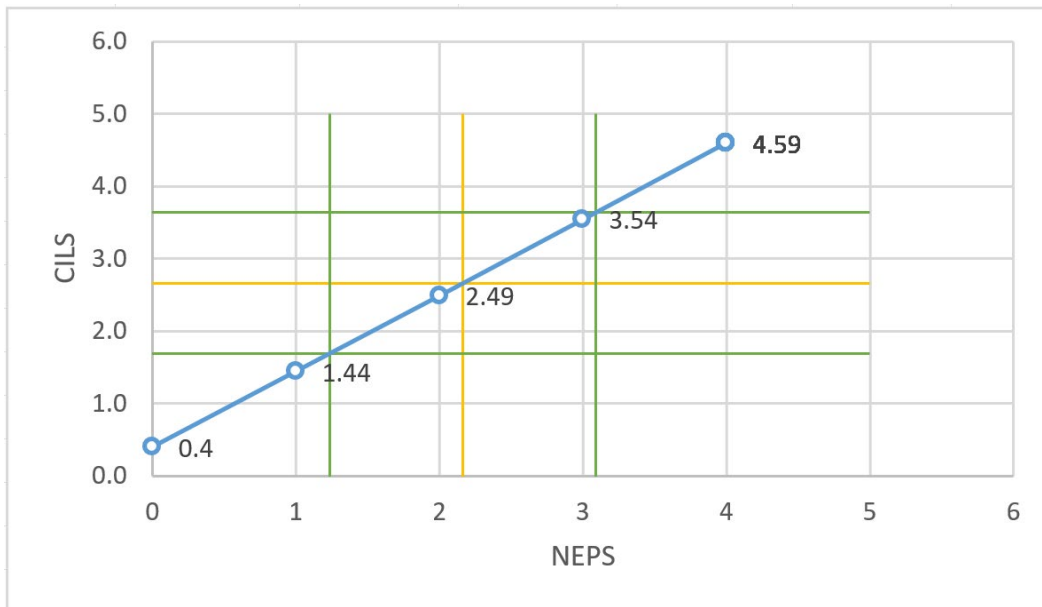


Abbildung 8: Grafische Darstellung der linearen Transformation und der Rekodierungswerte

Basierend auf der in der Excel-Tabelle erstellten Rekodierungstabelle werden die Antwortkategorien des NEPS SC4-Items in den Daten rekodiert. Aus dem Beispiel in Abbildung 8: Alle Beobachtungen mit einem Wert von „2“ erhalten bei der Umkodierung den Wert „2,49“, alle Beobachtungen mit einem Wert von „3“ erhalten bei der Umkodierung den Wert „3,54“, etc. Daraus ergibt sich eine Skala für das harmonisierte Item, die die CILS4EU-Antwortskala beibehält, aber mittlere Kategorien für die rekodierten NEPS SC4-Werte enthält. Die Bezeichnungen für die Antwortkategorien basieren ausschließlich auf den CILS4EU-Bezeichnungen (siehe Abbildung 10).

H_spm - H: Subjective school performance, Math

	Freq	Percent	Valid
-95 Implausible value	37	0.03	0.04
-90 Unspecific missing	236	0.16	0.28
-88 No nasser	148	0.10	0.18
-77 Not applicable	28	0.02	0.03
-66 Question not asked	850	0.59	1.02
-55 Other missing	13	0.01	0.02
-54 Missing by design	33642	23.30	40.18
-44 Interrupted Interview	3	0.00	0.00
1.00 Very well	6343	4.39	7.58
1.44	3270	2.26	3.91
2.00 Quite well	11676	8.09	13.95
2.49	6167	4.27	7.37
3.00 OK	10417	7.21	12.44
3.54	4385	3.04	5.24
4.00 Not that well	3687	2.55	4.40
4.59	1482	1.03	1.77
5.00 Not well at all	1341	0.93	1.60
Total	83725	57.99	100.00
Missing .	60659	42.01	
Total	144384	100.00	

Abbildung 9: Antwortskala eines Zielitems nach Linearem Equating

3.6. Harmonisierung der Leistungsmessungen

Die Harmonisierung der Reasoningstests basiert auf Summenscores der jeweiligen Leistungsmessungen, die in den Datensätzen der Schulbefragungen enthalten sind. In beiden Erhebungen wurden in Klasse 9 Kompetenzmessungen zur Erfassung der fluiden kognitiven Leistungsfähigkeit der Schüler:innen (Reasoning) durchgeführt. In CILS4EU wurden Items aus CFT20R ("Grundintelligenztest Skala 2 Revision", Weiß 2006; vgl. Abbildung 11) verwendet (CILS4EU, 2016), in NEPS SC4 der NEPS-MAT Test (siehe Haberkorn & Pohl, 2013; vgl. Abbildung 12).

Die Reasoning-Fähigkeiten der Jugendlichen wurden damit in beiden Studien auf sehr ähnliche Weise mittels Matrizen tests erfasst: Die logischen Regeln zur Anordnung geometrischer Elemente in 2x2- bzw. 3x3-Matrizen mussten erschlossen werden, um aus den angebotenen Lösungen jeweils ein Teil (bzw. bei manchen Aufgaben in CILS4EU zwei Teile) auszuwählen, mit denen die Darstellung korrekt vervollständigt wird. In CILS4EU standen je Aufgabe fünf, in NEPS SC4 sechs mögliche Lösungen zur Wahl. Die Gestaltung der Tests wurde bewusst so gewählt, dass sie möglichst muttersprachunabhängig veränderungsfrei über alle Altersgruppen hinweg effektiv eingesetzt werden können (Leibniz-Institut für Bildungsverläufe, 2018).

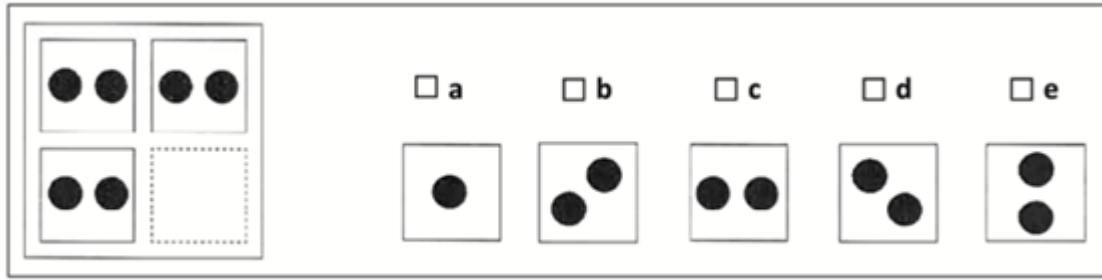


Abbildung 10: CILS4EU - Beispielitem aus dem Beispielitem aus CFTR20R (Quelle: CILS4EU (2016:42), ursprünglich aus Weiß, 2006)

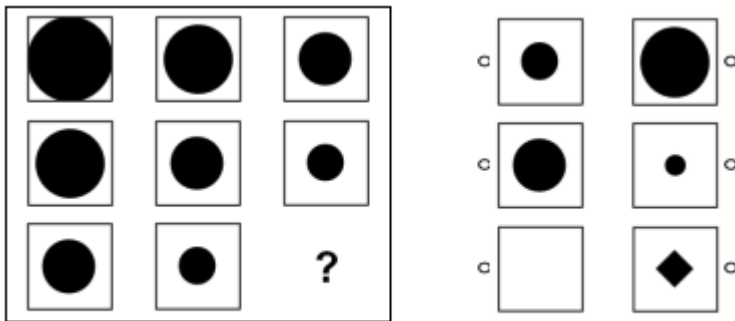


Abbildung 3: Beispielaufgabe des NEPS-MAT

Abbildung 11: NEPS SC4 - Beispielitem aus dem Reasoning-Test NEPS-MAT. (Quelle: Lang et al. (2014:10))

Für die Harmonisierung der Reasoning-Tests mussten Unterschiede in Konzeption und Durchführung berücksichtigt werden. In CILS4EU wurden 27 Matrizentests eingesetzt (Teil A mit 15, Teil B mit 12 Aufgaben), die ein breites Spektrum an Schwierigkeit abdecken – von sehr leicht bis anspruchsvoll – mit Erfolgsquoten zwischen 27 % und 99 %. Die Bearbeitungszeit betrug vier Minuten für Teil A und drei Minuten für Teil B. Da in NEPS SC4 kein Zeitdruck bestand, wurden Anpassungen vorgenommen, um dessen Auswirkungen in CILS4EU zu reduzieren. In NEPS SC4 umfassten die Reasoning-Tests 12 Aufgaben, gegliedert in drei Sets à vier Items. Innerhalb der Sets wurde die Schwierigkeit gesteigert, insgesamt lag sie jedoch enger beieinander (48 %–88 % richtige Lösungen). Je Set standen drei Minuten Bearbeitungszeit zur Verfügung.

Zur Prüfung der Vergleichbarkeit wurden Analysen zu Itemschwierigkeit, Trennschärfe, Skalenhomogenität, Nonresponse und Reliabilität durchgeführt. Auf Grundlage dieser Ergebnisse wurden alle 12 NEPS-Tests beibehalten, in CILS4EU jedoch nur 18 der 27 Aufgaben berücksichtigt. Die ersten vier Items (y1_cot1–y1_cot4) wurden ausgeschlossen, da sie von über 95 % der Jugendlichen richtig gelöst wurden und somit kaum Varianz aufwiesen. Ebenso wurden die letzten Aufgaben in Teil A (y1_cot13–y1_cot15) und Teil B (y1_cot26–y1_cot27) verworfen, da hier der Zeitdruck zu erhöhter Nonresponse und abnehmender Trennschärfe führte.

Aus den verbleibenden 18 Items in CILS4EU und den 12 Items in NEPS SC4 wurden Summenscores gebildet, wobei richtige Antworten mit 1, falsche und fehlende mit 0 kodiert wurden – entsprechend dem Vorgehen der Ursprungserhebungen. Da die Summenscores nicht

normalverteilt waren, wurde ein Equipercentile Equating (Kolen & Brennan, 2014) durchgeführt, um eine gemeinsame Skala zu berechnen, die den Vergleich der Reasoning-Fähigkeiten zwischen beiden Studien ermöglicht (Singh, 2021, 2022).

3.7. Umgang mit Klassifikations-Heuristiken in CILS4EU und NEPS SC4

Die Harmonisierung der Informationen zum Generationenstatus und zum Herkunftsland nimmt eine Sonderrolle ein, da die zugrunde liegenden Klassifikationsstrategien und Kodierungsregeln in CILS4EU und NEPS SC4 teils unterschiedlichen Heuristiken folgen (vgl. Dollmann et al., 2014; Olczyk et al., 2014). Ziel dieses Arbeitsschritts ist es, die Variablen zum Generationenstatus und zu den Herkunftsländern so zu harmonisieren, dass sowohl die CILS4EU- als auch die NEPS-SC4-Klassifikationslogik berücksichtigt wird. Insgesamt werden vier Variablen erstellt – jeweils zwei zum Generationenstatus und zwei zu den Herkunftsländern. Obwohl beide Datensätze einer ähnlichen Logik folgen, ist eine 1:1-Übertragung der jeweiligen Klassifikationsstrategien nicht möglich, da sich die Kodierungen der Herkunftsländer und die Definitionen der Missingcodes unterscheiden.

Bevor die NEPS-SC4-Klassifikation auf die CILS4EU-Daten angewendet werden konnte, wurden Informationen aus dem CILS4EU-Elterndatensatz (vgl. CILS4EU, 2014) mit dem aufbereiteten CILS4EU-Datensatz im langen Format verknüpft, um zusätzliche Angaben zu den Herkunftsländern der Großeltern zu integrieren. Diese Informationen liegen in der regulären CILS4EU-Download-Version nur in vereinfachter Form vor. Anschließend wurden äquivalente Variablen identifiziert und aufbereitet, wobei insbesondere auf unterschiedliche Werte- und Missingkodierungen geachtet wurde, um Informationsverluste oder Überschreibungen zu vermeiden.

Da CILS4EU neben Deutschland auch in drei weiteren Erhebungsländern durchgeführt wurde, musste bei der Anwendung des NEPS-SC4-Ansatzes eine Hilfsvariable erstellt werden, die zwischen im jeweiligen Befragungsland Geborenen und im Ausland Geborenen unterscheidet. Analog zum NEPS-SC4-Ansatz wurden zudem Variablen entwickelt, die fehlende oder widersprüchliche Informationen im Zuge der Erstellung der Zielvariablen identifizieren. Die so nach NEPS-SC4-Logik gebildeten Generationenstatus- und Herkunftslandvariablen auf Basis der CILS4EU-Daten stehen anschließend für die Harmonisierung zur Verfügung.

Umgekehrt wurde auch die CILS4EU-Klassifikationsstrategie auf die NEPS-SC4-Daten angewendet. Die dabei erstellten Variablen wurden in den harmonisierten Datensatz integriert und stehen für weiterführende Analysen zur Verfügung.

3.8. Erstellung des harmonisierten Datensatzes

Der harmonisierte Datensatz umfasst alle als unproblematisch bzw. als komplizierter klassifizierten Zielitems sowie die jeweils zugrunde liegenden Originalvariablen aus CILS4EU und NEPS SC4. Die Schüler:innenbefragung wird – basierend auf der Kombination beider Studien – im langen Datenformat bereitgestellt; die Datensätze der Elternbefragung, der COVID-19-Zusatzbefragung und der Reasoningtests liegen als separate Datensätze vor. Zur

eindeutigen Kennzeichnung erhalten die harmonisierten Variablen das Präfix „H_“ (Schüler:innen), „Hp_“ (Eltern) bzw. „Hc_“ (COVID-19); die Herkunft der Beobachtungen wird über „H_dtset“ ausgewiesen. Die ID-Variable „H_ID“ vereint die ursprünglichen Identifikatoren beider Studien.

Um in den Schüler:innenbefragungen die Wellen anzugeben, in welchen die jeweiligen Variablen gemessen wurden, wurden zwei Wellen-Indikatoren „H_wave“ und „H_wave_2“ erstellt. „H-wave“ enthält die ursprüngliche Wellenstruktur der NEPS SC4 mit acht Erhebungswellen, denen die fünf CILS4EU Wellen entsprechend ihrer zeitlichen Passung zugeordnet wurden. Da jedoch in CILS4EU für den Erhebungszeitraum nur fünf Wellen verfügbar sind, umfassen die Wellen „2“, „4“ und „6“ der „H_wave“-Variablen nur die NEPS SC4-Wellen. Die „H_wave“-Variable erlaubt Benutzer:innen, je nach Fragestellung freier zu entscheiden, welche NEPS SC4 Wellen mit den CILS4EU Wellen abgeglichen werden sollen. So ist es z.B. möglich, die beiden SchulabgängerInnenwellen (Welle 4 und 6 in NEPS SC4) mit Welle 2 bzw. Welle 3 in CILS4EU zu vergleichen. Zusätzlich zu „H_wave“ enthält der harmonisierte Datensatz die Wellen-Variable „H_wave_2“. Diese Variable umfasst fünf statt acht Wellen und gibt die beste zeitliche Passung der CILS4EU und NEPS SC4 Wellen an. Dies sind Welle 1, 3, 5, 7 und 8 des NEPS SC4. Abbildung 13 verdeutlicht die Unterschiede zwischen „H_wave_2“ und „H_wave“ in Form einer Tabelle.

3.9. Gewichtung

Durch die Kombination der beiden Datensätze müssen auch die vorhandenen Gewichtungen angepasst werden. Prinzipiell handelt es sich bei beiden Stichproben um zwei nicht-disjunkte Samples, allerdings wurden Schulen für die CILS4EU Stichprobe so ausgewählt, dass es zu keiner Überschneidung mit dem NEPS SC4 Sample kam. Eine Gewichtung des harmonisierten Datensatzes wurde durch die Arbeitseinheit Stichprobenziehung, Gewichtung und Imputation am LfBi erstellt. Sie sind dabei in Hinblick auf eine größtmögliche Effizienz der statistischen Analysen basierend auf dem harmonisierten Datensatz hin konstruiert. Da in CILS4EU in Welle 6 zusätzlich zu den Panelbefragten Befragte aus einer Auffrischungstichprobe hinzukamen, wurde ein weiteres Gewicht erstellt, das für die danach durchgeführte COVID-19 Zusatzbefragung verwendet werden kann.

3.10. Erstellung der Dokumentation

Zur Sicherstellung der Transparenz des Harmonisierungsprozesses – insbesondere bei der Erstellung und Klassifizierung der Zielitems – enthält der harmonisierte CILS4NEPS-Datensatz eine umfassende Dokumentation. Diese umfasst die Überblickstabelle, die Kodierungstabelle, den technischen Bericht zum Datensatz, das Codebuch, Stata-Do-Files zur Erstellung der Zielitems und des harmonisierten Datensatzes, Stata-Do-Files zu Herkunftsland und Generationenstatus, eine Excel-Vorlage zum Linear Equating sowie ein R-Skript zum Equipercentsile Equating der Leistungsmessungen.

Alle Dokumentationsmaterialien wurden sorgfältig geprüft und in ein einheitliches, benutzerfreundliches Format überführt. Überblicks- und Kodierungstabelle folgen dem in den

vorangehenden Kapiteln beschriebenen Aufbau. Für die Do-Files wurde ein Stata-Projekt angelegt, das den Nutzer:innen ein geordnetes Nachvollziehen der einzelnen Arbeitsschritte ermöglicht. Die Vorlage zum Linear Equating wurde in Excel auf Grundlage der einschlägigen Literatur entwickelt. Überblickstabelle, Kodierungstabelle und Linear-Equating-Vorlage enthalten jeweils ein begleitendes ReadMe, das Aufbau, Logik und Anwendung der Dokumente erläutert und so eine transparente Nachnutzung der Datenharmonisierung gewährleistet.

3.11. Datenarchivierung und -zugang

Die Daten aus CILS4EU sind über das GESIS-Datenarchiv zugänglich. Die Daten von NEPS SC4 werden durch das Forschungsdatenzentrum des LIfBi (FDZ-LIfBi) bereitgestellt. Es werden drei Möglichkeiten zum Zugang zu den NEPS-Daten angeboten: On-Site Gastarbeitsplätze mit dem höchsten Informationsgehalt der Daten, Datenfernzugriff (Remote-Access) für mittleren Informationsgehalt und ein Download der Daten mit geringerem Informationsgehalt und dem stärksten Anonymisierungsgrad. Um die ausschließlich über Remote-Access verfügbaren Informationen aus NEPS SC4 verwenden zu können, werden die harmonisierten CILS4NEPS Daten nach Beantragung über das RemoteNEPS Gateway zur Verfügung gestellt. Der Zugang zu den NEPS-Daten, RemoteNEPS und CILS4EU-Daten kann separat beantragt werden, die harmonisierten CILS4NEPS-Daten können jedoch nur über RemoteNEPS genutzt werden. Das harmonisierte CILS4NEPS Datenprodukt wird auf der CILS4EU-Homepage (www.cils4.eu) und der NEPS-Data-Seite (www.neps-data.de/Datenzentrum/Daten-und-Dokumentation/SC4-Klasse-9/CILS4NEPS) gelistet.

Der Zugang zu den CILS4NEPS Daten am FDZ des LIfBi wird nur dann bewilligt, wenn aus dem Antrag ersichtlich ist, dass tatsächlich eine Forschungsfrage verfolgt wird, die mit Hilfe des harmonisierten Datensatzes beantwortet werden soll. Andernfalls wird die antragstellende Person an GESIS verwiesen, um dort – wie bisher – CILS4EU zu beantragen. In der RemoteNEPS-Umgebung werden außerdem Skriptdateien zur Verfügung gestellt, anhand derer die vorhandene Harmonisierung repliziert und bei Bedarf geändert oder erweitert werden kann.

Die Prüfung der Anträge erfolgt in Kooperation zwischen dem FDZ des LIfBi und dem Projektteam von CILS4EU, das derzeit die Anträge für CILS4EU vor der Datenfreigabe bearbeitet.

4. Zusammenfassung und Fazit

Mit der Harmonisierung von CILS4EU und NEPS SC4 wurde ein einzigartiges Datenangebot geschaffen, das die Untersuchung von Bildungs- und Integrationsprozessen in Deutschland in bislang nicht erreichter Detailtiefe ermöglicht. Durch die Zusammenführung beider Datensätze entstehen erweiterte Analysemöglichkeiten – insbesondere für Gruppen- und Übergangsanalysen, die mit einem Einzeldatensatz nur eingeschränkt möglich wären. Zugleich wird es möglich, deutsche Befunde in die internationale Vergleichsperspektive von CILS4EU einzubetten.

Das Projekt hat gezeigt, dass Ex-post-Harmonisierung weit mehr ist als eine rein technische Aufgabe: Sie erfordert konzeptionelle Entscheidungen über Äquivalenzen, Messlogiken und Konstruktdefinitionen. Durch die Kombination aus systematischer Dokumentation, methodischer Reflexion und Bereitstellung aller notwendigen Materialien wurde durch das Harmonisierungsprojekt ein Vorgehen etabliert, das Nachvollziehbarkeit, Transparenz und Replizierbarkeit gewährleistet.

Mit dem harmonisierten CILS4NEPS-Datensatz steht der Forschung künftig ein leistungsfähiges Instrument zur Verfügung, das neue empirische Analysen über Bildungswege, Integration und ethnische sowie soziale Ungleichheiten ermöglicht. Zugleich leistet das Projekt einen Beitrag zur Weiterentwicklung methodischer Standards der Datenharmonisierung und zur besseren Nutzung bestehender sozialwissenschaftlicher Infrastrukturen in Deutschland und Europa.

Perspektivisch ist vorgesehen, den harmonisierten Datensatz um ein Record Linkage mit IAB-Daten zu ergänzen – in Analogie zu NEPS-ADIAB (Bachbauer & Wolf, 2023) und CILS4EU-DE-ADIAB (Antoni et al., 2025). Die Verknüpfung erfolgt – wie in den bestehenden Projekten – auf einem etwas stärker aggregierten Datenbestand. Dadurch gehen zwar Detailstufen verloren, zugleich eröffnen sich aber neue Auswertungsmöglichkeiten durch die Anbindung an die Integrierten Erwerbsbiografien (IEB) der IAB-Daten.

Literaturverzeichnis

- Antoni, M., Dollmann, J., Schmucker, A., & Weißmann, M. (2025). Children of Immigrants Longitudinal Survey in Four European Countries – Germany verknüpft mit administrativen Daten des IAB (CILS4EU-DE-ADIAB) 1975-2023. (*FDZ-Datenreport 05/2025* (de)), S. 86. <https://doi.org/10.5164/IAB.FDZD.2505.de.v1>
- Bachbauer, N.; & Wolf, C. (2023). NEPS-SC4-Erhebungsdaten verknüpft mit administrativen Daten des IAB (NEPS-SC4-ADIAB). (*FDZ-Datenreport 10/2023* (de)), Nürnberg, S. 96. <https://doi.org/10.5164/IAB.FDZD.2310.de.v1>
- Blossfeld, H.-P., Rossbach, H.-G., & von Maurice, J. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, 14 (Sonderheft), 19–34.
- CILS4EU. (2016). *Children of immigrants longitudinal survey in four European countries: Technical report. Wave 1 – 2010/2011 (Version 1.2.0)*. Mannheim University.
- Dollmann, J., Arnold, L., Horr, A., Kerzner, V., Schmidt, R., Soiné, H., Weber, F. & Weißmann, M. (2023). Technical Report: A Harmonised Dataset Based on CILS4EU and NEPS SC4 (CILS4NEPS), Version 1.0. Mannheim: Mannheim University.
- Dollmann, J., Jacob, K., & Kalter, F. (2014). *Examining the diversity of youth in Europe: A classification of generations and ethnic origins using CILS4EU data* (MZES Working Paper No. 156). Mannheim Centre for European Social Research.
- Dollmann, J., Arnold, L., & Horr, A. (2024). CILS4NEPS – Unlocking research potential through more participants, more schools and international comparison: Harmonized data for research on education, school-to-work transition and integration processes for adolescents in Germany, the Netherlands, Sweden and England. *Journal of Economics and Statistics*, 245(1-2), 215–234.
- Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 315–334). Wiley.
- Haberkorn, K., & Pohl, S. (2013). *Cognitive basic skills – Data in the Scientific Use File*. National Educational Panel Study, University of Bamberg.
- Hoffmeyer-Zlotnik, J. H. (2008). Harmonisation of demographic and socio-economic variables in cross-national survey research. *Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique*, 98, 5–24.
- Kalter, F., Heath, A. F., Hewstone, M., Jonsson, J. O., Kalmijn, M., Kogan, I., & van Tubergen, F. (2016). *Children of immigrants longitudinal survey in four European countries (CILS4EU): Full version*.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.
- Lang, F. R., et al. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie* (NEPS Working Paper No. 43). Leibniz Institute for Educational Trajectories.

- Leibniz-Institut für Bildungsverläufe. (2018). *Informationen zur Kompetenztestung: NEPS Startkohorte 4—Klasse 9. 2. Welle: 9. Jahrgangsstufe (Studie A47 Haupterhebung 2011/12)*.
https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/4-0-0/NEPS_SC4_Competerences_w2_de.pdf
- Nusser, L., & Messingschlager, M. (2018). *Erfassung kognitiver Grundfähigkeiten bei Schülerinnen und Schülern an Förderschulen in Startkohorte 4 (Klasse 9)* (NEPS Survey Paper No. 33). Leibniz Institute for Educational Trajectories.
<https://doi.org/10.5157/NEPS:SP33:1.0>
- Olczyk, M., Will, G., & Kristen, C. (2014). *Immigrants in the NEPS: Identifying generation status and group of origin* (NEPS Working Paper No. 41a). Leibniz Institute for Educational Trajectories.
- Schneider, S. L. (2016). *Die Konzeptualisierung, Erhebung und Kodierung von Bildung in nationalen und internationalen Umfragen*. GESIS – Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_020-1
- Singh, R. K. (2020). Harmonizing instruments with equating. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, 6, 11–18.
- Singh, R. K. (2021). Harmonizing data in the social sciences with equating. In T. Wolbring, H. Leitgöb, & F. Faulbaum (Eds.), *Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter* (pp. 123–140). Springer VS.
- Singh, R. K. (2022). Harmonizing single-question instruments for latent constructs with equating using political interest as an example. *Survey Research Methods*, 16(3), Article 3. <https://doi.org/10.18148/srm/2022.v16i3.7916>
- StataCorp. (2021). *Stata statistical software: Release 17*. StataCorp LLC.
- Weiß, R. H. (2006). *Grundintelligenztest Skala 2 (CFT 20)*. Hogrefe.
- Wolf, C., Schneider, S. L., Behr, D., & Joye, D. (2016). Harmonizing survey questions between cultures and over time. In C. Wolf, D. Joye, T. Smith, & Y.-C. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 502–524). SAGE.

Impressum

Kontakt:

Lena Arnold

Mannheimer Zentrum für Europäische Sozialforschung MZES

E-Mail: lena.arnold@uni-mannheim.de

Mannheim, Juni 2026

KonsortSWD Working Paper:

KonsortSWD baut als Teil der Nationalen Forschungsdateninfrastruktur Angebote zur Unterstützung von Forschung mit Daten in den Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften aus. Unsere Mission ist es, die Forschungsdateninfrastruktur zur Beforschung der Gesellschaft zu stärken, zu erweitern und zu vertiefen. Sie soll nutzungsorientiert ausgestaltet sein und die Bedürfnisse der Forschungscommunities berücksichtigen.

Wichtiger Grundstein ist dabei das seit über zwei Jahrzehnten durch den Rat für Sozial- und Wirtschaftsdaten (RatSWD) aufgebaute Netzwerk von Forschungsdatenzentren.

In dieser Reihe erscheinen Beiträge rund um das Forschungsdatenmanagement, die im Kontext von KonsortSWD entstehen. Beiträge, die extern und doppelblind begutachtet wurden sind entsprechend gekennzeichnet.

KonsortSWD wird im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert – Projektnummer: 442494171.



Diese Veröffentlichung ist unter der Creative-Commons-Lizenz (CC BY 4.0) lizenziert:

<https://creativecommons.org/licenses/by/4.0/>

DOI: 10.5281/zenodo.1912908

Zitationsvorschlag:

Arnold, L., Horr, A., Bauer, M., Finke, B., & Dollmann, J. (2026). *Harmonisierung und Zusammenführung der Daten des Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) mit den Daten der Startkohorte 4 des Nationalen Bildungspanel (NEPS) – CILS4NEPS*. KonsortSWD Working Paper 16/2026. Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften (KonsortSWD – NFDI4Society).

<https://doi.org/10.5281/zenodo.1912908>