Bericht zum Forschungsvorhaben

„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 -

im Rahmen der Projektförderung Forschungsdatenmanagement durch das Konsortium für die Sozial-, Verhaltens-, Bildungs-, und Wirtschaftswissenschaften (KonsortSWD)

Antragsteller:

Jörg Dollmann

Mannheimer Zentrum für Europäische Sozialforschung MZES

Universität Mannheim

Postfach

68131 Mannheim

Andreas Horr

Leibniz-Institut für Bildungsverläufe

Bildungsentscheidungen und -prozesse, Migration, Bildungsrenditen

Wilhelmplatz 3

96047 Bamberg

Berichtszeitraum: 01.07.2021–30.06.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Zielsetzung und Hintergrund des Projekts

1.1 Allgemeine Zielsetzung

Ziel des Harmonisierungsprojekts war es, zusätzliche Forschungspotentiale durch die Kombination der beiden Datenquellen „Children of Immigrants Longitudinal Survey in Four European Countries“ (CILS4EU; Kalter *et al.*, 2016) und „Startkohorte 4 des Nationalen Bildungspanels (NEPS)“ (nachfolgend NEPS SC4 bzw. NEPS; Blossfeld, Rossbach and Maurice, 2011) zu erschließen, die mit den jeweils einzelnen Datensätze nicht oder nur bedingt möglich wären. 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.

1.2 Beschreibung und Vergleichbarkeit der Datenkorpora

CILS4EU

CILS4EU ist eine internationale Längsschnittstudie, deren Ziel es ist, die Integration von Jugendlichen mit und ohne Migrationshintergrund in Deutschland, England, den Niederlanden und Schweden zu untersuchen. Das Studiendesign orientiert sich an dem der NEPS SC4. Neben der expliziten Stratifizierung mit einem Oversampling migrantenreicher Schulen wurde in den einzelnen Ländern zusätzlich implizit stratifiziert; in Deutschland nach Bundesland und Schulart, um diese Merkmale proportional zur Grundgesamtheit zu berücksichtigen.

Insgesamt wurden in der ersten Welle ca. 19.000 Jugendliche befragt (5.000 in Deutschland), wobei etwa die Hälfte einen Migrationshintergrund besitzt. Im Rahmen der internationalen Förderung wurden zwischen 2011 und 2013 zwei weitere Erhebungswellen realisiert, wobei die zweite Welle ebenfalls im Schulkontext stattfand, während die TeilnehmerInnen der dritten Welle außerhalb des Schulkontextes online, postalisch oder telefonisch befragt wurden. Nach der dritten Welle wurde der deutsche Teil von CILS4EU in das Langfristprogramm der DFG

aufgenommen. Im Jahr 2016, im Zuge der sechsten Erhebungswelle, wurde eine Auffrischungsstichprobe gezogen, in der ebenfalls Jugendliche bzw. junge Erwachsene mit Migrationshintergrund überproportional häufig vertreten waren. Aktuell sind die Daten aus acht Wellen (zzgl. einer Welle zur COVID-19 Pandemie) verfügbar, die Erhebung der neunten Welle startete im ersten Quartal 2022.

NEPS SC4

Das NEPS erhebt Längsschnittdaten zu Kompetenzentwicklungen, Bildungsprozessen, Bildungsentscheidungen und Bildungsrenditen in formalen, nicht-formalen und informellen Kontexten. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom BMBF finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.

Die Teilstudie SC4 untersucht, beginnend mit Jahrgangsstufe 9, Wege in und durch die Sekundarstufe II ebenso wie Übergänge in das berufliche Bildungssystem, in ein Studium sowie in den Arbeitsmarkt. Zielpopulation stellt die Schülerschaft an Regelschulen und Förderschulen dar, die im Schuljahr 2010/11 Klasse 9 besuchten. Dazu wurde eine geschichtete Klumpenstichprobe bei Regelschulen sowie eine Stichprobe mit Jugendlichen an Förderschulen gezogen.

Die Jugendlichen waren zum Zeitpunkt der ersten Erhebung 14–15 Jahre alt, insgesamt liegen für die erste Welle für etwas mehr als 15.500 Jugendliche Befragungsdaten vor, von denen etwa 37% einen Migrationshintergrund besitzen. Jugendliche, die weiterhin die ausgewählten Schulen besuchten, wurden in den folgenden Wellen im Schulkontext befragt. Schulabgänger wurden außerhalb des Schulkontextes (überwiegend über CATI) verfolgt.

Die Haupterhebung an den Schulen wurde als PAPI durch IEA-DPC durchgeführt, die CATI und CAWI-Befragungen im Individualfeld durch infas – Institut für angewandte Sozialwissenschaften. Aktuell sind für SC4 Daten aus zwölf Wellen (zzgl. einer Welle zur COVID-19-Pandemie) verfügbar.

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 2010/11) und CILS4EU einen sehr ähnlichen Samplingansatz auf Schulebene wie NEPS SC4 umgesetzt hat, dabei aber Schulen mit hohen Migrantenanteilen überproportional häufig gezogen wurden. 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.

Im Folgenden soll dargelegt werden, wie die beiden Datenquellen zunächst dahingehend evaluiert wurden, wo sie tatsächliche oder potenzielle inhaltliche Überschneidungen aufweisen. Hierzu wurden die Fragen und Antwortoptionen der beiden Datenquellen miteinander verglichen (s.h. Abschnitt 2.1 und 2.2). Im Anschluss erfolgte die theoretische sowie empirische Erstellung der Zielitems und des harmonisierten CILS4NEPS Datensatzes (s.h. Abschnitt 2.3 bis 2.7). Abschließend berichten wir über Gewichte im harmonisierten Datensatz sowie die Erstellung der Dokumentation und Zugang zum harmonisierten Datensatz (s.h. Abschnitt 2.8 bis 2.10).

2 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. Im Folgenden wird zunächst das Ziel des jeweiligen Arbeitsschrittes verdeutlicht. Anschließend werden die einzelnen Unter-Arbeitsschritte erläutert. Insgesamt werden für die Harmonisierung ausschließlich die Zielpersonen als Befragtengruppe verwendet, nicht aber Informationen aus zusätzlichen Fragebögen für andere Personengruppen, wie beispielsweise Eltern oder LehrerInnen.

Ex-post Harmonisierung erklärt

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, Wolf and Hadorn, 2010). Im Fall der CILS4NEPS Harmonisierung ist die Ex-post Harmonisierung die verfügbare Strategie, 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.

Insgesamt gesprochen existieren keine fest etablierten Schritte für die Ex-post Harmonisierung von Daten, jedoch schlagen Experten folgenden Schritte vor (Granda, Wolf and Hadorn, 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.

2.1 Überblickserstellung

Ziel dieses Arbeitsschrittes

In einem ersten Schritt der Harmonisierung von Daten aus CILS4EU und NEPS SC4 wurden beide Datenquellen auf ihre inhaltlichen Gemeinsamkeiten hin untersucht, um ihr Harmonisierungspotenzial zu ermitteln. Diese Überblickstabelle enthält Informationen zu allen in CILS4EU Welle 1–3 enthaltenen Variablen sowie zu potenziell äquivalenten Variablen aus NEPS SC4. Abbildung 1 zeigt einen exemplarischen Ausschnitt aus der Überblickstabelle.

Die Tabelle ist dabei wie folgt aufgebaut: Die rechte Seite der Tabelle enthält alle CILS4EU Variablen, die linke Seite die potenziell entsprechenden Variablen aus dem NEPS SC4. Die Variablen sind mit ihrem Variablennamen, ihrer Fragestellung, sowie den Antwortkategorien in den Zeilen eingetragen. Dabei wird sowohl für CILS4EU wie auch NEPS SC4 ersichtlich, für welche Befragtengruppe in welcher Welle die jeweilige zugrundeliegende Information erhoben wurde.

CILS4EU (w1-w3)

Bereich	Indicator	Variable name	Question	Categories	Students Wave 1	Parents Wave 1	Student Wave 2	Student 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 Startkohorte 4 = 9.Klasse (w1-w10)

			Schüler																	
			Welle 1		Welle 2		Welle 3		Welle 4	Welle 5				Welle 6	Welle 7					
							Absolventen	Schüler	Absolventen	Schüler		Schulabgänger		Schulabgänger	Schüler	Individuelle Nachverfolgung		Absolventen	
			Regelschulen	Förderschulen	Regelschulen	Förderschule		Regelschulen								Klasse 11	Klasse 12		
Variable Name	Frage	Kategorien	Erstbefrag	Erstbegfragte	Erstbefragte	Erstbegfragte	Erstbefragte	Erstbefragte	Panelbefragte	Erstbefragte	Panelbefragte	Erstbefragte	Panelbefragte	Panelbefragte		Erstbefragte	Panelbefragte		
Socio-Demograph																			
Geschlecht																			
t700031	Bist du... Sind Sie...	...männlich? ...weiblich?	x	x	x	x	x (t700001)	x	x (t700001)	x	x	x (t00001)	x (t700001)	x (t700001)	x	x	.	x (t700001)	x (t700001)

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

Die Tabelle dient dem Zweck einer ersten grundlegenden Evaluation, inwieweit eine Übereinstimmung der erhobenen Informationen bzw. der Variablen im CILS4EU und NEPS SC4 vorliegt.

Um die Äquivalenz der Variablen zu klassifizieren, wurde ein Farbschema erstellt; wobei grün hinterlegte Variablen (nahezu) identisch hinsichtlich der Fragestellung und Antwortkategorien sind, gelb hinterlegte Variablen ähnliche Konstrukte erfassen, aber nicht identisch sind und rot hinterlegte Variablen nur in einem der Datensätze existieren. Anhand einer Zählung der grünen und gelben Variablen in der Überblickstabelle lässt sich somit direkt bestimmen, wie viele Variablen Potenzial für eine Harmonisierung mitbringen. Die Überblickstabelle bildet damit eine wichtige Grundlage für alle weiteren Arbeiten zur tatsächlichen Harmonisierung der Variablen. Auch lassen sich in dieser Überblickstabelle direkt die Variablennamen und notwendigen Befragtengruppen erkennen, welche relevant sind für die Zusammenführung der Variablen von CILS4EU und NEPS SC4 in sogenannte Zielitems. Als Zielitems werden dabei diejenigen Variablen im harmonisierten Datensatz bezeichnet, die aus der Kombination der CILS4EU und NEPS SC4 Variablen entstehen. Darüber hinaus stellt diese Überblickstabelle ein Verzeichnis aller für die Harmonisierung denkbaren Variablen dar, das der Dokumentation, der Transparenz und Nachvollziehbarkeit dienen soll. Die Überblickstabelle ist diesem Bericht als Anhang beigefügt.

Arbeitsschritte

Im ersten Schritt der Erstellung der Überblickstabelle wurden aus allen drei CILS4EU Wellen (Wellen 1 bis 3) enthaltenen Variablen aufgelistet, jeweils mit der Angabe, in welchen Wellen und für welche Personengruppen sie abgefragt wurden. Die Variablen wurden gemäß dem CILS4EU Codebuch Inhaltsbereichen zugeordnet, die später zur Harmonisierung einzeln bearbeitet werden können.

In einem zweiten Schritt wurden für jede CILS4EU Variable inhaltliche Äquivalente aus dem NEPS SC4 gesucht unter Verwendung der NEPS SC4 SUF-Fragebögen und des Codebuchs. Diese ermittelten NEPS SC4 Variablen wurden ebenfalls mit Variablenname und Antwortkategorie eingetragen. Zusätzlich wurden die entsprechenden NEPS SC4 Wellen, wie folgt gekennzeichnet:

- mit einem X, wenn eine CILS4EU äquivalente Frage gestellt wurde,
- einem X(...), wenn die Frage gestellt wurde, aber innerhalb der NEPS SC4 Wellen einen abweichenden Variablennamen hatte,

- einem Punkt, wenn die Variable im NEPS SC4 nicht in annähernd ähnlicher Form enthalten war.

Nachdem alle Variablen eingegeben waren, konnte somit ermittelt werden, welche Variablen im CILS4EU und NEPS SC4 identisch (grün hinterlegt), welche ähnlich, aber nicht identisch (gelb hinterlegt) und welche CILS4EU Variablen nicht in ähnlicher Form im NEPS SC4 enthalten waren (rot hinterlegt). Insgesamt kommen auf Basis dieses Überblicks circa 260 Variablen für die Harmonisierung grundsätzlich in Frage.

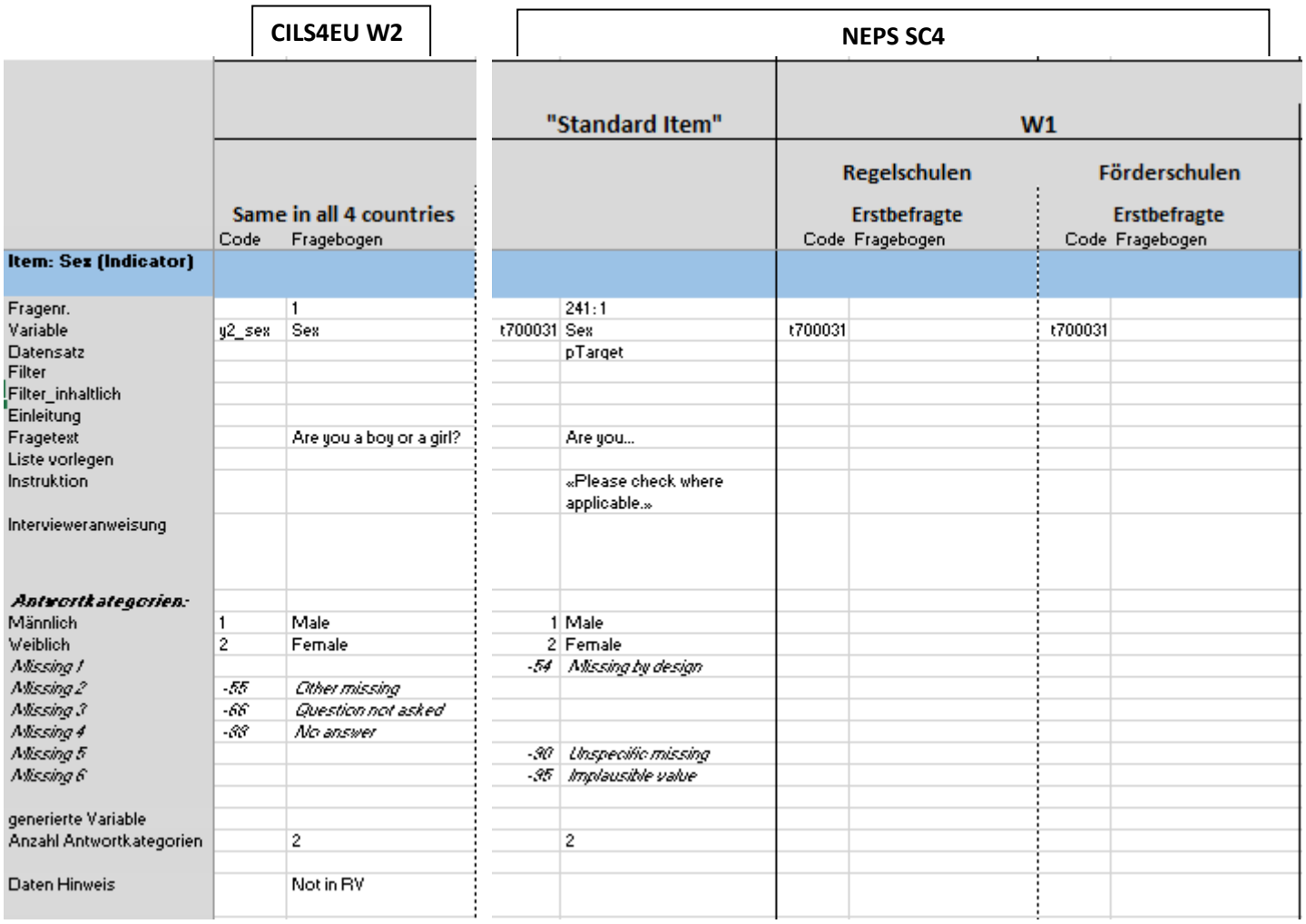
2.2 Kodierungsüberblick

Ziel dieses Arbeitsschrittes

Zweck der Kodierungsübersicht in Form einer Excel-Tabelle ist es, eine präzise Übersicht zu erhalten, wie die zu harmonisierenden Variablen aus CILS4EU und NEPS SC4 kodiert sind, um diese Kodierung dann entsprechend bei der Harmonisierung der Variablen im Softwarepaket Stata (StataCorp., 2021) zu berücksichtigen. Die Kodierungstabelle gliedert sich dabei in verschiedene Excel-Blätter, die den inhaltlichen Bereichen der Variablen entsprechen, zum Beispiel „Allgemeine Informationen“ oder „Haushaltssituation“. Diese Kodierungstabelle ist der Grundstein für die Harmonisierung der Daten in Stata. Ohne jeweils die Codebücher oder Fragebögen wiederholt heranziehen zu müssen, kann die komplette Harmonisierungsarbeit in den Daten auf dieser Tabelle beruhen. Sie dokumentiert dabei auch genau die Harmonisierung der Variablen und bildet somit das Grundstück der Datenharmonisierung. Die Kodierungstabelle liegt dem Bericht als Anhang bei.

Arbeitsschritte

Die Kodierungstabelle wurde wie folgt erstellt. Ausgehend von der Überblickstabelle werden die zu harmonisierenden Variablen aus CILS4EU und NEPS SC4 in den entsprechenden, jeweils nach Inhaltsbereichen geordneten, Excel-Blättern der Kodierungstabelle eingetragen. Hierzu werden gleiche Informationen und gleiche Kodierungen stets in derselben Zeile gelistet. Die folgende Abbildung veranschaulicht die Eintragung einer Variable in die Kodierungstabelle.



	CILS4EU W2		NEPS SC4					
			"Standard Item"		W1			
					Regelschulen		Förderschulen	
	Same in all 4 countries				Erstbefragte		Erstbefragte	
	Code	Fragebogen			Code	Fragebogen	Code	Fragebogen
Item: Sex (Indicator)								
Fragenr.		1		241:1				
Variable	y2_sex	Sex	t700031	Sex	t700031		t700031	
Datensatz				pTarget				
Filter								
Filter_inhaltlich								
Einleitung								
Fragetext		Are you a boy or a girl?		Are you...				
Liste vorlegen								
Instruktion				«Please check where applicable.»				
Intervieweranweisung								
Antwortkategorien:								
Männlich	1	Male	1	Male				
Weiblich	2	Female	2	Female				
Missing 1			*-54*	*Missing by design*				
Missing 2	*-55*	*Other missing*						
Missing 3	*-66*	*Question not asked*						
Missing 4	*-88*	*No answer*						
Missing 5			*-90*	*Unspecific missing*				
Missing 6			*-95*	*Implausible value*				
generierte Variable								
Anzahl Antwortkategorien		2		2				
Daten Hinweis		Not in RV						

Abbildung 2. Auszug aus der Kodierungstabelle

Wie Abbildung 2 zu entnehmen ist, ist für die CILS4EU und NEPS SC4 Variablen dokumentiert, welche Antwortkategorien auf welche Weise kodiert werden, ebenso für die Kodierung der „missings“. Für die NEPS SC4 Variablen ist zudem kenntlich gemacht, in welchem Datensatz die Variable abgelegt ist, d.h. welcher NEPS SC4 Datensatz später zur Harmonisierung herangezogen werden muss. Zentral ist hierbei, dass gleiche Antwortkategorien (unabhängig von ihrer Kodierung) in die gleiche Zeile eingetragen werden. Somit werden Übereinstimmungen in den Antwortkategorien sowie Abweichung in den Kodierungen zwischen gleichen Antwortkategorien unmittelbar ersichtlich. Im obigen Beispiel ist die Variable für Geschlecht in CILS4EU und NEPS SC4 exakt gleich kodiert, sodass in diesem Fall keine Re-Kodierung für die Harmonisierung in den Daten erforderlich ist.

Dieses Verfahren wird zudem auch für die jeweiligen „missings“ angewendet. Auf Grund der spezifischen Kodierungen und Klassifizierungen der „missings“ in CILS4EU und NEPS SC4

werden diese jedoch zunächst untereinander aufgelistet und am Ende der Harmonisierung in ein einheitliches missing-Schema überführt. Zusätzlich werden Filter-Fragen in der Kodierungstabelle ergänzt, da diese gegebenenfalls zu harmonisieren sind. Variablen die in der Überblickstabelle als rot (in CILS4EU aber nicht in NEPS SC4 vorhanden) markiert wurden, sind nicht in der Kodierungstabelle eingetragen.

2.3 Theoretische Erstellung der Zielitems

Ziel dieses Arbeitsschrittes

Ziel dieses Arbeitsschrittes ist es, eine exakte Arbeitsvorlage für die zu generierenden harmonisierten Zielitems in den Daten zu erstellen und die Erzeugung sowie deren Zusammensetzung in der Kodierungstabelle zu dokumentieren. Als Zielitems werden diejenigen Variablen im harmonisierten Datensatz bezeichnet, die aus der Kombination der CILS4EU und NEPS SC4 Variablen entstehen. Somit lässt sich anhand der in der Kodierungstabelle eingetragenen Zielitems erkennen, wie die jeweiligen Variablen aus CILS4EU und NEPS SC4 rekodiert und zusammengeführt werden müssen – die Ausgangslage für die Erstellung der Zielitems in Stata.

Arbeitsschritte

Zunächst werden Fragenummer und Variablenname für das Zielitem vergeben und es wird dokumentiert, aus welchen CILS4EU und NEPS SC4 Variablen das betreffende Zielitem zu erstellen ist. Anschließend wird die Vergleichbarkeit der Variablen bewertet – sowohl hinsichtlich des gemessenen Konstrukts, wie auch ihrer Antwortkategorien. Für die Beurteilung dieser Vergleichbarkeit haben wir uns an einschlägiger Forschungsliteratur zur Ex-post Harmonisierung von Umfragedaten orientiert (z.B. Wolf *et al.*, 2016; Granda, Wolf and Hadorn, 2010; Hoffmeyer-Zlotnik, 2008). Des Weiteren standen und stehen uns Dr. Verena Ortmanns und Dr. Ranjit Singh vom GESIS beratend bei der Harmonisierung – insbesondere bei der Beurteilung der Zielitems und deren Harmonisierungsstrategien – zur Seite.

Hierbei wird geprüft, ob die zugehörigen Variablen aus CILS4EU und NEPS SC4 dasselbe Konstrukt in Hinblick auf die Fragestellung messen und ob es sich bei diesem Konstrukt um ein beobachtbares (z.B. Bildungsstand, Anzahl der Personen im Haushalt) oder ein latentes (nicht-beobachtbares) Konstrukt handelt. In einem weiteren Schritt werden die Antwortkategorien auf ihre Ähnlichkeit (inhaltliche Ähnlichkeit sowie Ähnlichkeit in Bezug auf Form und Anzahl der Antwortkategorien) hin untersucht. Diese Beurteilung ist zentral für die später in den Daten zur Anwendung kommende Harmonisierungsstrategie. Die

Vergleichbarkeit wird mittels eines Farbschemas klassifiziert, das auch im harmonisierten Datensatz und Codebuch ausgewiesen sein wird. Während in der Kodierungsübersicht jedoch alle klassifizierten Variablen ausgewiesen sind, enthält die erste Version des CIL4NEPS Datensatzes nur als *unproblematisch* und als *komplizierter* klassifizierte Variablen.

Unproblematisch-klassifizierte Zielitems beruhen auf Variablen, die in den jeweiligen Datensätzen als beobachtbare Konstrukte gemessen werden. Als Beispiel kann hier „Geburtsmonat“ aufgeführt werden, welches sowohl im CILS4EU wie auch im NEPS SC4 als „Wann sind Sie geboren, Monat?“ abgefragt und mit Zahlen von 1 „Januar“ bis 12 „Dezember“ kodiert wird (vgl. Abbildung 3). Durch diese Übereinstimmung in Konstrukt-Ähnlichkeit und identischen oder sehr ähnlichen Antwortkategorien, kann für dieses Zielitem ein Harmonisierungsverfahren in Form einer einfachen Zuordnung (und ggf. Re-Kodierung) der Antwortkategorien durchgeführt werden.

Item: Day of birth, Month	**Item: Day of Birth - Mont**	**CILS4EU**		**NEPS**	
2	Question number		2		241 : 2
H_dobm Day of birth, Month	Variable	y3_dobm	Day of Birth, Month	t70004m	Day 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?
	Submit list				
	Instruction				
	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*
-98 Don't know	*Missing 9*			*-98*	*Don't know*
-99 Filtered	*Missing 10*			*-99*	*Filtered*
	Generated variable				
12	Number of response categories		12		12
	Data remark				

Abbildung 3. Beispiel eines als *unproblematisch*-klassifizierten Zielitems – „Geburtsmonat“

Komplizierter-klassifizierte Zielitems messen in beiden Datensätzen latente Konstrukte und sind hinsichtlich des gemessenen Konstrukts und den Antwortkategorien identisch bis ähnlich.

Es kann sich auch um manifeste (beobachtbare) Konstrukte handeln, die mehr als eine einfache Zuordnung der Antwortkategorien erfordern (z.B. durch ‚lagging' der Antwortkategorien). Ein Beispiel für ein latentes Konstrukt ist „Selbstwirksamkeit in der Schule" (vgl. Abbildung 4). Hier wird ein ähnliches Konstrukt in CILS4EU und NEPS SC4 gemessen. Die Antwortkategorien zwischen CILS4EU und NEPS SC4 unterscheiden sich in geringfügiger Weise (Zustimmung auf einer 5-Punkt Skala vs. Zutreffen auf einer 4-Punkt Skala). Für diese Zielitems würde eine einfache Zuordnung der Antwortkategorien zu Verzerrungen in Analysen führen. Diese Zielitems können zwar von allen künftigen DatennutzerInnen verwendet werden, es wird jedoch empfohlen, diese Items in den Analysen zu validieren. Beispielsweise empfiehlt es sich zu prüfen, ob die harmonisierten Zielitems mit bestimmten theoretisch erwartbaren Variablen korrelieren (Singh, 2021; Singh, 2020; Kolen and Brennan, 2014). Als Harmonisierungsstrategie wenden wir Linear Equating an, das wir in Abschnitt 2.5 genauer erläutern.

Item: Self-efficacy - do well at school		Item: Self-efficacy I		CILS4EU		NEPS
54		Question number		16		740 : 55214i
H_sesch	**Self-efficacy - do well at school**	Variable	y2_seff1	Self-efficacy I	t66002c	Self-concept school: good in most school subjects
		Data record				pTarget
	CILS: y*_seff1 NEPS: t66002c					
		Filter				
		Filter_contentual				
		Introduction				
	Agreement/ disagreement: I am good at school.	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				
		Antwortkategorien:				
Rational numbers	Strongly agree / Does completely apply	Strongly agree / Does completely apply	1	Strongly agree	4	does completely apply
Rational numbers	.	Agree / Does rather apply	2	Agree	3	does rather apply
Rational numbers	.	Neither agree nor disagree	3	Neither agree nor disagree		
Rational numbers	.	Disagree / Does rather not apply	4	Disagree	2	does rather not apply
Rational numbers	Strongly disagree / Does not apply at all	Strongly disagree / Does not apply at all	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
		Data remark				

Abbildung 4. 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) verdeutlich wird, messen die Fragen zwar ähnliche Konstrukte, jedoch werden im CILS4EU die Antwortkategorien in Form von Häufigkeiten und im NEPS SC4 in Form von Zustimmung gemessen. Diese Zielitems werden als „semi-problematisch“ klassifiziert. Sie werden in dieser ersten Version des CILS4NEPS, die innerhalb der Konsort SWD Förderperiode liegt, nicht harmonisiert. Im weiteren Verlauf der CILS4NEPS Harmonisierung über den Förderzeitraum hinaus, werden als *semi-problematisch-klassifizierte Zielitems* hinsichtlich ihrer Vergleichbarkeit tiefer gehend evaluiert und wenn möglich harmonisiert.

Problematisch-klassifizierte Zielitems werden von uns gekennzeichnet, jedoch nicht harmonisiert. Obwohl es in beiden Datensätzen Variablen gibt, die in einem rudimentären Sinne ein verwandtes Konstrukt messen könnten, bestehen doch zu viele Abweichungen (in Frage- und/oder Antwortkategorien), so dass die Konstruktvergleichbarkeit nicht mehr gegeben ist.

Ein Beispiel hierfür (vgl. Abbildung 6) ist „Geschlechterrollen“. Dieses Item wird im CILS4EU als *single-barrel* item gefragt, mit Antwortkategorien gegliedert nach Mann/Frau/Beiden. Im NEPS SC4 wird das Item mit doppelter Bedeutung abgefragt mit den Antwortkategorien in Form von Zustimmung. Durch diese Abweichungen sind starke Verzerrungen im Zielitem und in weiteren Analysen zu erwarten, weshalb dieses Zielitem nicht in den Daten erstellt 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 5. Beispiel eines als semi-problematisch-klassifizierten Zielitems (das Zielitem wurde nicht in die Kodierungstabelle eingetragen)

			CILS4EU				*NEPS*	
Attitudes and Norms	Gender roles: Child care	grol1	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 6. Beispiel eines als problematisch-klassifizierten Zielitems (das Zielitem wurde nicht in die Kodierungstabelle eingetragen)

2.4 Datenaufbereitung CILS4EU und NEPS SC4

Ziel dieses Arbeitsschrittes

Ziel dieses Arbeitsschrittes ist es, die Datensätze von CILS4EU und NEPS SC4 in ein einheitliches Format zu bringen, um darauffolgend mit der Harmonisierung in dem zusammengeführten Datensatz von CILS4EU und NEPS SC4 beginnen zu können.

Arbeitsschritte

Die CILS4EU Daten werden in einem „weiten" Datenformat bereitgestellt, bei dem jede befragte Person einer Zeile in den Daten entspricht. Um Konvergenz mit dem NEPS SC4 Datensatz herzustellen, der in einem langen Datenformat (Beobachtungen pro befragte Person in mehreren Zeilen) bereitgestellt wird, wurden die ersten drei Wellen von CILS4EU kombiniert und ebenfalls in ein langes Datenformat konvertiert. Aufgrund der Struktur des weiten Datenformats enthält jede Variable im originalen CILS4EU Datensatz ein Präfix (z.B. „y1_'), das angibt, in welcher Welle die jeweilige Variable erhoben wurde. Auch wenn die Items über die Erhebungswellen gleich gehalten werden, kommt es vereinzelt zu kleinen Abweichungen bei den Antwortkategorien. Daher waren bei der Konvertierung der CILS4EU Daten in ein langes Datenformat Änderungen in der Beschriftung der Antwortkategorien für bestimmte Variable erforderlich. Wir dokumentieren diese Änderungen an den Original-Daten im Anhang des Technischen Berichts für CILS4NEPS. Um anzuzeigen, welche Variable aus dem CILS4EU Datensatz stammen, haben wir allen diesen Variablen das Präfix „CILS4EU_" vorangestellt.

Für die NEPS SC4 Daten waren keine Änderungen in der Struktur der Daten erforderlich. Wir identifizierten die für die Harmonisierung relevanten Datensätze: „Cohort Profile", „pTarget", „pTargetCATI" und „Weights" und kombinierten diese gemäß der vom NEPS Datenzentrum bereitgestellten Merging-Matrix. Wie beim CILS4EU Datensatz fügten wir allen Variablen aus dem NEPS das Präfix „NEPS_" hinzu, bevor wir beide Datensätze zusammenführten.

2.5 Erstellung der Zielitems in den Daten

Ziel dieses Arbeitsschrittes

Nachdem die Zielitems in theoretischer Weise in der Kodierungstabelle erstellt wurden, erläutert dieser Arbeitsschritt wie diese Items nachfolgend in den Daten erstellt wurden.

Arbeitsschritte

Um die harmonisierten Variablen in den Daten zu erzeugen, entscheiden wir zunächst über die anzuwendende Harmonisierungsstrategie. Die Entscheidung dieser Strategie hängt mit der oben eingeführten Klassifizierung in *unproblematische* und *kompliziertere* Items zusammen. Während wir bei den *unproblematischen* Zielitems eine einfache Zuordnung der Antwortkategorien angewandt haben, wurde bei *komplizierteren* Zielitems das Verfahren des Linearen Equatings verwendet. Beide Harmonisierungsstrategien werden im Folgenden knapp beschrieben.

Die *Zuordnung der Antwortkategorien* beider Quell-Variablen erforderte in Teilen eine umgekehrte Kodierung der Antwortskalen oder eine Zusammenfassung mehrerer Antwortkategorien zu einer Kategorie – je nach Kodierung der Quell-Variablen. In anderen Fällen wurden mehrere Variablen aus CILS4EU oder NEPS SC4 zu einer harmonisierten Variable zusammengefasst. Ein Beispiel hierfür ist die Variable „Haushaltsmitglieder", für die im NEPS SC4 verschiedene Arten von potenziellen Haushaltsmitgliedern (z. B. Mutter, Stiefmutter, Adoptivmutter) in einer einzigen Frage abgefragt werden, während in CILS4EU diese Informationen in separaten Fragen bewertet werden. Daher basiert das harmonisierte Zielitem „Haushaltsmitglieder – Mutter, Stiefmutter, Adoptivmutter" auf einer NEPS SC4 Variablen, aber 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; Singh, 2021). Wichtig für die Anwendung dieses Verfahrens ist, dass die zu harmonisierenden Instrumente aus derselben Grundgesamtheit gezogen wurden (Kolen and 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 and 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 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 7. 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 (s.h. 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 8. 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“).

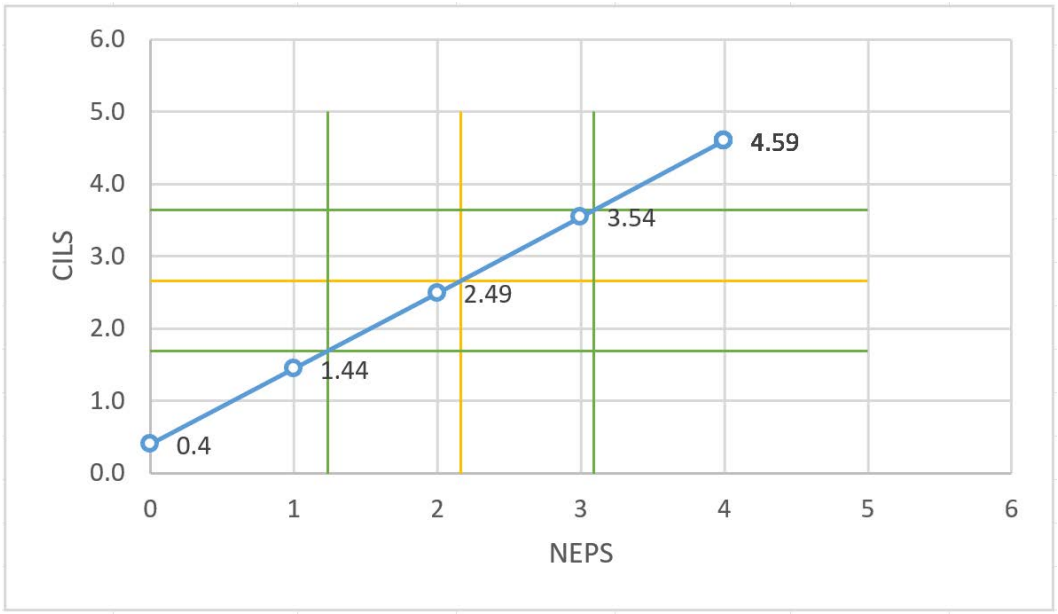


Abbildung 9. Grafische Darstellung der linearen Transformation und der Rekodierungswerte

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	Cum.
Valid	-95	Implausible value	37	0.03	0.04	0.04
	-90	Unspecific missing	236	0.16	0.28	0.33
	-88	No answer	148	0.10	0.18	0.50
	-77	Not applicable	28	0.02	0.03	0.54
	-66	Question not asked	850	0.59	1.02	1.55
	-55	Other missing	13	0.01	0.02	1.57
	-54	Missing by design	33642	23.30	40.18	41.75
	-44	Interrupted Interview	3	0.00	0.00	41.75
	1	Very well	6343	4.39	7.58	49.33
	1.440000057220458984		3270	2.26	3.91	53.23
	2	Quite well	11676	8.09	13.95	67.18
	2.490000009536743164		6167	4.27	7.37	74.55
	3	OK	10417	7.21	12.44	86.99
	3.539999961853027344		4385	3.04	5.24	92.22
	4	Not that well	3687	2.55	4.40	96.63
	4.590000152587890625		1482	1.03	1.77	98.40
	5	Not well at all	1341	0.93	1.60	100.00
	Total		83725	57.99	100.00	
Missing	.		60659	42.01		
Total			144384	100.00		

Abbildung 10. Antwortskala eines Zielitems nach Linearem Equating

2.6 Umgang mit Klassifikations-Heuristiken im CILS4EU und NEPS SC4

Der Harmonisierung der Informationen zum *Generationenstatus* und *Herkunftsland* kommt eine Sonderrolle zu, da die Klassifikationsstrategien je nach Datensatz teils unterschiedlichen Heuristiken sowie Kodierungsregeln bei der Definition dieser Konstrukte folgen (vgl. Dollmann, Jacob and Kalter, 2014; Olczyk, Will and Kristen, 2014).

Ziel dieses Arbeitsschrittes

Ziel ist es, die Variablen zum Generationenstatus sowie den Herkunftsländern unter Berücksichtigung i) der CILS4EU Klassifikationsstrategie sowie auch ii) des NEPS SC4 Ansatzes zur Klassifikation des Generationenstatus und der Herkunftsländer zu harmonisieren.

Somit werden insgesamt vier Variablen zur Verfügung gestellt mit jeweils zwei Generationenstatus- und Herkunftslandvariablen. Wenngleich die zugrundliegenden Heuristiken in beiden Datensätzen einer vergleichbaren Logik folgen, kann generell keine 1:1 Anwendung der CILS4EU Klassifikationsstrategie auf Basis der NEPS SC4 Daten und umgekehrt erfolgen. Ursächlich hierfür sind unter anderem variierende Wertekodierungen der Herkunftsländer sowie unterschiedliche Definitionen der Missingcodes je nach Datensatz.

Arbeitsschritte

i) Bevor die NEPS SC4 Klassifikation des Generationenstatus auf den CILS4EU Datensatz Anwendung finden konnte, wurden Informationen aus dem CILS4EU-Elterndatensatz (vgl. CILS4EU, 2014) zum aufbereiteten CILS4EU Datensatz im langen Format hinzugespielt, um detaillierte Informationen über Herkunftsländer der Großeltern der Befragten zu nutzen, da diese in der on-site Version der CILS4EU Daten für Großeltern in einer vereinfachten Form zur Verfügung stehen. Zusätzlich wurden äquivalente CILS4EU Variablen identifiziert und aufbereitet. Wichtig ist hierbei die Prüfung der CILS4EU Variablen auf potenziell unterschiedliche Werte- und Missingkodierungen, um eine Überschreibung von Informationen auszuschließen

Darüber hinaus ist zu berücksichtigen, dass im CILS4EU Datensatz neben Deutschland drei weitere Befragungsländer (*survey country*) enthalten sind und somit bei der Klassifizierung des Generationenstatus nach dem NEPS SC4 Ansatz in die Kodierung zu integrieren sind. Hierzu wurde eine Hilfsvariable erstellt, welche die Identifikation der im jeweiligen Befragungsland Geborenen von den nicht im jeweiligen Befragungsland Geborenen ermöglicht. Die expliziten Kodierungs-Schritte sind dem technischen Bericht des CILS4NEPS zu entnehmen.

Analog zum NEPS SC4 Ansatz wurden drei weitere Variablen erstellt, die fehlende bzw. widersprüchliche Informationen im Rahmen der Erstellung der oben genannten Zielvariablen identifizieren. Die Generationenstatus und Herkunftslandvariablen nach der NEPS SC4 Klassifikationsstrategie sind auf Basis der CILS4EU Daten erstellt worden und stehen für die weitere Harmonisierung zur Verfügung.

ii) Die CILS4EU Klassifikationsstrategie zur Erstellung der Generationenstatus- und Herkunftslandvariablen wurde umgekehrt auf Basis der NEPS SC4 Daten erstellt. Die erstellten Variablen sind dem harmonisierten Datensatz hinzugespielt und stehen zur weiteren Nutzung bereit.

2.7 Erstellung des harmonisierten Datensatzes

Ziel dieses Arbeitsschrittes

Der zu erstellende harmonisierte Datensatz enthält als *unproblematisch* oder als *komplizierter* klassifizierte Zielitems sowie die den jeweiligen Zielitems zugrunde liegenden Originalvariablen aus CILS4EU und NEPS SC4.

Arbeitsschritte

Basierend auf der Kombination von CILS4EU and NEPS SC4 wird der harmonisierte Datensatz in einem langen Datenformat zur Verfügung gestellt. Alle harmonisierten Zielitems erhalten das Präfix „H_“ sowohl in ihrem Variablennamen als auch in der Variablenbeschreibung. Die Variable „H_dtset“ gibt an aus welchen Datensatz die jeweiligen Beobachtungen stammen. Als ID-Variable dient „H_ID“, welche die Ursprungswerte aus den ID Variablen von CILS4EU („youthid“) und NEPS SC4 („ID_t“) enthält.

Um 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 des NEPS SC4 mit sechs Erhebungswellen, denen die drei CILS4EU Wellen entsprechend ihrer zeitlichen Passung zugeordnet wurden. Da jedoch in CILS4EU für den Erhebungszeitraum nur drei Wellen verfügbar sind, umfassen die Wellen „2“, „4“ und „6“ der „H_wave“-Variablen nur die NEPS SC4 Wellen. Auch wenn insgesamt nur drei Wellen für Panelanalysen mit den harmonisierten Daten verwendet werden können, erlaubt die „H_wave“-Variable den BenutzerInnen, je nach Fragestellung freier zu entscheiden, welche NEPS SC4 Wellen mit den CILS4EU Wellen abgeglichen werden sollen. So ist es möglich, die beiden SchulabgängerInnenwellen (Welle 4 und 6 im 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 drei statt sechs Wellen und gibt die beste zeitliche Passung der CILS4EU und NEPS SC4 Wellen an. Dies sind Welle 1, 3 und 5 des NEPS SC4. Abbildung 11 verdeutlicht die Unterschiede zwischen „H_wave_2“ und „H_wave“ in Form einer Kreuztabelle.

H_wave2

H_wave	H: Harmonized wave 1-6	H: Harmonized wave 1-3				
		Wave only	Wave 1	Wave 2	Wave 3	Total
H_wave	Wave 1	0	35,323	0	0	35,323
	Wave 2 (NEPS only)	16,425	0	0	0	16,425
	Wave 3	0	0	32,215	0	32,215
	Wave 4 (NEPS only)	16,425	0	0	0	16,425
	Wave 5	0	0	0	27,571	27,571
	Wave 6 (NEPS only)	16,425	0	0	0	16,425
	Total	49,275	35,323	32,215	27,571	144,384

Abbildung 11. Kreuztabelle der harmonisierten Wellen-Variablen „H_wave_2“ und „H_wave“

2.8 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. Dennoch müssen die jeweiligen Designgewichte nach einer Kombination der beiden Samples angepasst werden. Für die Erstellung der Gewichte kooperieren wir mit der Arbeitseinheit Stichprobenziehung, Gewichtung und Imputation am LIfBi. Die kombinierten Gewichte sind dabei in Hinblick auf eine größtmögliche Effizienz der statistischen Analysen basierend auf dem harmonisierten Datensatz hin konstruiert.

2.9 Erstellung der Dokumentation

Ziel dieses Arbeitsschrittes

Um den Harmonisierungsprozess – insbesondere die Erstellung und Klassifizierung der Zielitems – für die Daten UserInnen transparent zu gestalten, wird dem harmonisierten

CILS4NEPS Datensatz eine ausführliche Dokumentation beigefügt. Diese umfasst die folgenden Komponenten:

1) Überblickstabelle
2) Kodierungstabelle
3) Technischer Bericht zum Datensatz
4) Codebuch
5) Do-Files zur Erstellung der Zielitems und des harmonisierten Datensatzes
6) Do-Files zu Herkunftsland und Generationenstatus
7) Excel-Vorlage zu Linear Equating

Arbeitsschritte

Alle als Dokumentation bereitgestellten Materialen wurden und werden von uns gegengeprüft und in ein einheitliches sowie verständliches Format gebracht. Die Überblickstabelle und die Kodierungstabelle folgen dabei dem oben unter den jeweiligen Überschriften beschriebenen Aufbau. Für die Do-Files wurde ein Stata-Projekt angelegt, das den NutzerInnen ein geordnetes Nachvollziehen der einzelnen Kodierungsschritte erlaubt. Die Vorlage zu Linear Equating wurden in Excel, auf Basis der oben zitierten Literatur erstellt. Die Überblickstabelle, die Kodierungstabelle sowie die Linear Equating Vorlage erhalten jeweils ein „ReadMe“, in welchem die Logik und Lesart des jeweiligen Dokuments für NutzerInnen erläutert wird.

2.10 Datenarchivierung und -zugang

Die Daten von NEPS SC4 werden durch das Forschungsdatenzentrum des LIfBi (FDZ-LIfBi) bereitgestellt. Die Daten aus CILS4EU sind über das GESIS-Datenarchiv zugänglich. 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 dabei zusätzlich zur CILS4EU Homepage (www.cils4.eu) perspektivisch auch auf der neps-data Seite (www.neps-data.de/Datenzentrum/Daten-und-Dokumentation) gelistet – wodurch Sichtbarkeit und Bekanntheit der CILS4NEPS Daten erhöht werden.

Der Zugang zu den CILS4NEPS Daten am FDZ-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. Im Sinne der NutzerInnenfreundlichkeit umfasst das Datenangebot jedoch den gesamten CILS4EU Datensatz (Welle 1 bis 3) und zusätzlich die harmonisierten Variablen. Zusätzlich werden in der RemoteNEPS-Umgebung Skriptdateien zur Verfügung gestellt, anhand derer die vorhandene Harmonisierung repliziert und bei Bedarf geändert oder erweitert werden kann (s.h. Dokumentation).

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

2.11 Workshop zum harmonisierten CILS4NEPS Datenprodukt

Im Rahmen des CILS4NEPS Datenprodukts sind zwei Arten von Workshops in Planung. Der erste Workshop ist für den für den Herbst 2022 angesetzt. In diesem Workshop werden die harmonisierten CILS4NEPS Daten einer ausgewählten Personengruppe vorgestellt und Feedback eingeholt. Diese Personengruppe soll auf der einen Seite Experten zur Ex-post Harmonisierung von Umfragedaten umfassen, wie beispielsweise Dr. Silke Schneider, Dr. Verena Ortmanns und Dr. Ranjit Singh. Auf der anderen Seite, werden ausgewählte Personen, die sowohl die CILS4EU wie auch die NEPS SC4 Daten zu eigenen Forschungszwecken verwenden eingeladen. Dieser Workshop hat ein eintägiges Format und wird aus Kostengründen, wie auch aus Unwägbarkeiten aufgrund der weiterhin angespannten COVID-19-Situation online stattfinden.

Der zweite Workshop findet unter dem Format des „Meet the Data“ statt. Hierbei sollen die Potentiale der harmonisierten CILS4NEPS Daten einer breiten NutzerInnenschaft präsentiert und Analysepotenziale aufzeigen werden. Auch dieser Workshop ist als eintägiges Format geplant und kann, entsprechend der Nachfrage, wiederholt werden. Er dient maßgeblich zur Promotion des CILS4NEPS Datenproduktes und somit zur Generierung einer breiten Daten-Nutzendenzahl. Dieser Workshop kann perspektivisch als Grundlage für ein Vertiefungsmodul in NEPS-Datenschulungen dienen.

3 Literatur

Blossfeld, H.-P., Rossbach, H.-G. and Maurice, J. von (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft,* **14**.

CILS4EU (2014). Children of Immigrants Longitudinal Survey in Four European Countries. Technical Report. Wave 1 – 2010/2011, v1.1.0.

Dollmann, J., Jacob, K. and Kalter, F. (2014). Examining the diversity of youth in Europe: A Classification of Generations and Ethnic Origins Using CILS4EU Data (Technical Report). *MZES Arbeitspapiere - Working Papers,* **156**.

Granda, P., Wolf, C. and Hadorn, R. (2010). Harmonizing Survey Data. In Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B.-E. and Smith, T. W. (Eds.). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts.* Hoboken, New Jersey: Wiley, pp. 315–334.

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. and van Tubergen, F. (2016). Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU) – Full version.

Kolen, M. J. and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking.* New York, NY: Springer New York.

Olczyk, M., Will, G. and Kristen, C. (2014). Immigrants in the NEPS: Identifying Generation Status and Group of Origin. *NEPS Working Paper,* **41a**.

Singh, R. K. (2020/2021). *Adventures in ex-post harmonization:* GESIS - Leibniz-Institut für Sozialwissenschaften.

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 Wolbring, T., Leitgöb, H. and Faulbaum, F. (Eds.). *Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter. Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute.* Wiesbaden: Springer VS, pp. 123–140.

StataCorp. (2021). *Stata Statistical Software: Release 17:* College Station, TX: StataCorp LLC.

Wolf, C., Schneider, S. L., Behr, D. and Joye, D. (2016). Harmonizing survey questions between cultures and over time. In Wolf, C., Joye, D., Smith, T. and Fu, Y.-c. (Eds.). *The SAGE Handbook of Survey Methodology.* London: SAGE Publications Ltd, pp. 502–524.

4 Anhang

- 1_Überblickstabelle
- 2_Kodierungstabelle