

gesis

Leibniz-Institut
für Sozialwissenschaften

Servicedokument

ALLBUS

Allgemeine Bevölkerungsumfrage
der Sozialwissenschaften

Kodierung und Definition von
fehlenden Werten im ALLBUS

Ein vereinheitlichtes Missing-Schema

Sonja Schulz

GESIS Datenarchiv für Sozialwissenschaften

GESIS - Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8
50667 Köln

Telefon: +49/(0)221/47694-0
Fax: +49/(0)221/47694-199
info@gesis.org, www.gesis.org

Inhalt

1	Kodierung und Definition von fehlenden Werten	2
1.1	Anforderungen an ein vereinheitlichtes Missing-Schema.....	2
2	Regeln zur Vergabe von fehlenden Werten	4
2.1	Aus dem Frage-Antwort-Prozess entstandene Missings.....	4
2.2	Kategorien, die aus inhaltlichen Erwägungen als fehlend deklariert werden – „Inhaltlicher Trifft-nicht-zu-Code“	5
2.3	Fehlende Angaben, weil einem Teil der Befragten die entsprechende Frage nicht gestellt wurde („Filter-Missings“).....	7
2.4	Missings nach Datenprüfung und -bereinigung.....	8
2.5	Technisches Missing.....	9
2.6	Missings bei generierten und abgeleiteten Variablen	9
2.6.1	Missings bei Kategorisierungen und gruppierten Variablen.....	9
2.6.2	Missings bei Kodierungen aus offenen oder halboffenen Fragen	9
2.6.3	Missings bei komplexeren Variablenkonstruktionen.....	11
2.7	Nicht enthaltene Informationen	12
2.7.1	Informationen, die bewusst nicht bereitgestellt werden	12
2.7.2	Zusätzliches Missing für ALLBUS-Kumulation	13

1 Kodierung und Definition von fehlenden Werten

In sozialwissenschaftlichen Umfragedaten finden sich in aller Regel fehlende Angaben und andere Merkmalsausprägungen, die im Sinne der meisten Anwendungen nicht sinnvoll interpretiert werden können und daher von der Analyse ausgeschlossen werden. Beispiele hierfür sind Antwortverweigerungen, Meinungslosigkeit, fehlende Antworten aufgrund von Filterführungen im Fragebogen, aber auch Fragen, die den Befragten gestellt wurden, obwohl diese den Befragten nicht betreffen und eine gültige Antwort im Sinne der Frageintention nicht möglich ist. Für diese Fälle werden bei den entsprechenden Variablen spezielle Codes mit zugehörigem Label zugewiesen und diese im Datensatz als fehlende Werte definiert (Statistikprogramm SPSS).

Für die Erhebungen des ALLBUS wurde eine einheitliche und sparsame Verwendung von fehlenden Werten (Missing Values) und den zugehörigen Labels beschlossen. Das neue Missing-Schema soll erstmalig für Datensatz und Dokumentation des ALLBUS 2016 umgesetzt werden.

Für die zukünftige Definition von fehlenden Werten in den ALLBUS-Daten sollen folgende Regelungen gelten:

- SPSS: Datensätze werden mit vordefinierten fehlenden Werten veröffentlicht.
- Stata: Um einheitliche Codes im Stata- und im SPSS-Datensatz zu gewährleisten, werden Stata-Datensätze ohne definierte Missing Values veröffentlicht, das heißt, die Werte, die in SPSS als fehlend definiert wurden, repräsentieren in den Stata-Datensätze gültige Zahlenwerte mit entsprechenden Kategorien-Labels. Zu jedem Stata-Datensatz wird ein Do-File angeboten, mit dem Nutzer die fehlenden Werte nach einem festen Schema auf „a“, „b“ usw. setzen können. Der Do-File wird vom ALLBUS-Nutzerservice bereitgestellt.

1.1 Anforderungen an ein vereinheitlichtes Missing-Schema

Werden bei der Zuweisung von fehlenden Werten bestimmte Konventionen beachtet, so erleichtert dies die Datenaufbereitung und -auswertung. Hierfür ist es zielführend, wenn ein Kategorienschema die folgenden Anforderungen erfüllt:

1. *Die Zuordnung von Codes für fehlende Werte soll möglichst eineindeutig sein.*

Dementsprechend gibt es für einen Ausfallgrund genau einen Missing-Code mit zugehörigem Label; im Umkehrschluss geht aus einem Missing-Code mit Label der Ausfallgrund eindeutig hervor. Ein Zusammenziehen von verschiedenen Ausfallgründen im gleichen Missing-Code soll vermieden werden. Hierdurch erhalten Nutzer genauere Informationen über den Grund, warum ein Fall keinen gültigen Wert aufweist. Weitergehende Variablenkonstruktionen werden hierdurch erleichtert (z.B. wenn Informationen zu einem bestimmten Merkmal aufgrund der Filterführung im Fragebogen in der interessierenden Variable nicht enthalten sind, aber als Antworten zu anderen Fragen abgelegt sind). Außerdem werden Nonresponse-Analysen vereinfacht, wenn diese das eigentliche Interesse der Nutzer darstellen (wenn bspw. Antwortverweigerungen analysiert werden sollen, wird die Analyse dadurch vereinfacht, wenn es für diese nur genau einen Code gibt, der nicht mit anderen Ausfallgründen gemeinsam kodiert wird).

2. *Fehlende Werte sollen möglichst konstant gelabelt werden.*

Dies ermöglicht im Prozess der Datenaufbereitung eine automatisierte Vergabe von Missing-Value-Codes und Labels und reduziert damit die Fehleranfälligkeit. Auf mögliche Abweichun-

gen von Belabelung der Antwortkategorien, die dem Befragten (im Selbstausfüllerformat) oder dem Interviewer präsentiert wurden, wird an geeigneter Stelle (z.B. im Variable Report) hingewiesen. Dies gilt grundsätzlich nicht nur für die Label von fehlenden Werten, sondern auch für andere Variablen-Labels, die aufgrund der gebotenen Kürze häufig sehr knapp umschrieben oder anderweitig trunziert werden.

3. *Missings sollen nicht auf Werte mit Nachkommastellen kodiert werden.*

Um die ALLBUS-Datensätze einheitlich und gleichermaßen fehlerfrei in Datenformaten für verschiedene Statistikprogramme wie SPSS und Stata vertreiben zu können, ist es erforderlich, Codes für fehlende Werte so zu wählen, dass diese in den im Vertrieb befindlichen Datenformaten gleichermaßen gelabelt werden können. Da in Stata Werten mit Nachkommastellen kein Label zugeordnet werden kann, sollten Codes für fehlende Werte ganzzahlig gewählt werden.

4. *Fehlende Werte sollen auf Zahlen im negativen Wertebereich kodiert werden.*

Für die ALLBUS-Datensätze war es bislang üblich, fehlende Werte auf positive Werte zu kodieren. Dies hat Nachteile in der Datenaufbereitung und -analyse, da der gültige Wertebereich einer Variablen die Anzahl an Stellen der Codes für fehlende Werte bestimmt. Hat eine Variable gültige Werte im Bereich 1 bis 5, wurden fehlende Werte beispielsweise auf 7, 8, 9 und 0 gesetzt. Nimmt eine Variable die Werte 0 bis 10 an, ist dies nicht möglich und ein mindestens zweistelliger Code für fehlende Werte muss gewählt werden. Dies führt dazu, dass demselben Ausfallgrund (z.B. einer Verweigerung), eine Vielzahl an verschiedenen Codes zugeordnet wird (z.B. 7, 97, 997, 9997, 9999.97). Dies verhindert sowohl ein einfaches Durchzählen von fehlenden Werten eines Befragten (und damit beispielsweise einfache Analysen zur ‚Ernsthaftigkeit‘ oder dem ‚betriebenen Aufwand‘ bei der Fragebeantwortung) als auch ein schnelles Umkodieren auf andere Werte für den gesamten Datensatz (z.B. auf systemdefiniert fehlende Werte in Stata). Das Kodieren auf fixe negative Werte ist eine Lösung für diese Probleme – jedoch nur, wenn der negative Wertebereich für fehlende Werte reserviert bleibt. Wurden im Befragungsinstrument Skalen mit negativen Skalenwerten versehen und dem Befragten entsprechend präsentiert, sollten diese auf den positiven Wertebereich rekodiert werden und ein negativer Zahlenwert in das Variablenlabel aufgenommen werden.

2 Regeln zur Vergabe von fehlenden Werten

Die zukünftige Vergabe von fehlenden Werten soll an den Punkten 1-4 orientiert erfolgen und entsprechend die Zuordnung von fehlenden Werten möglichst eindeutig vornehmen, fehlende Werte konstant belabeln und ganzzahlige Codes im negativen Wertebereich wählen. Der negative Wertebereich bleibt entsprechend für fehlende Werte reserviert. Neben der Anforderung der Eindeutigkeit von Missing-Codes ist jedoch zu beachten, dass kein ‚Wildwuchs‘ von Missing-Codes und Labels entsteht (z.B. durch verschiedene Filterstufen im Fragebogen, die jeweils einen eigenen Code mit eigenem Label erhalten, wie etwa „Trifft nicht zu, kein Partner im Haushalt“, „Trifft nicht zu, Partner nicht erwerbstätig“), sondern ein sparsames Kategorienschema gewählt wird. Die Anforderung der Eindeutigkeit kann im Einzelfall in einem Spannungsverhältnis zur Sparsamkeit stehen. Im Folgenden wurden daher verschiedene Ausfallsprozesse als Oberkategorien identifiziert, die sich in verschiedene einzelne Ausfallgründe untergliedern lassen. Hierzu ist festzustellen, dass sich diese in einigen Fällen möglicherweise grundsätzlich weiter untergliedern ließen. Gleichsam ließe sich eine größere Sparsamkeit durch das Zusammenfassung auf den Oberkategorien erreichen. Das gewählte Missing-Schema ist somit als ein Kompromiss zu verstehen, der versucht Eindeutigkeit und Sparsamkeit gleichermaßen zu optimieren.

Unter den Missing-Codes der verschiedenen Oberkategorien lässt sich zumeist ein Code identifizieren, der der Oberkategorie am Ähnlichsten ist, bzw. den größten Allgemeinheitsgrad hat. Wenn das Ziel ist, beispielsweise das Missing-Schema des ALLBUS mit dem anderer Umfrageprogramme kompatibel zu machen bzw. die Kategoriensysteme verschiedener Umfrageprogramme zu vereinheitlichen, kann es unter Umständen sinnvoll sein, die verschiedenen Missing-Codes einer Oberkategorie auf einen gemeinsamen Code umzukodieren. In der folgenden Darstellung der verschiedenen Missing-Codes nach Ausfallprozessen sind die jeweiligen ‚allgemeinsten Codes‘ durch Fettdruck hervorgehoben.

2.1 Aus dem Frage-Antwort-Prozess entstandene Missings

Den Missing Codes in dieser Oberkategorie ist gemein, dass einem Befragten eine Frage zur Beantwortung vorgelegt wurde, aber der Befragte keine Antwort auf diese Frage geben konnte oder wollte.

Code	Label	Erläuterung zur Vergabe
-6	Item nicht bekannt	Befragter kennt fragliches Objekt/Konstrukt nicht. Code wird auch vergeben, wenn Antwortkategorie bspw. lautete „kenne Partei nicht“. Wenn dieser Code vergeben wird, hat Interviewer/Befragter diese Antwortkategorie zusätzlich zu „weiß nicht/kann n. sagen“ erhalten. Beispiele: „Beteiligung an Bürgerinitiative“ wurde erfragt, und „kenne ich nicht“ als Antwortoption zusätzlich zu „weiß nicht“ vorgegeben. Wahrscheinlichkeit CDU zu wählen wurde erfragt und „kenne Partei nicht“ neben „weiß nicht“ angeboten.
-7	Explizite Verweigerung	Befragter gibt an, Frage nicht beantworten zu wollen. Vergabe dieses Codes setzt voraus, dass Befragter im Selbstausfüllerformat eine entsprechende Antwortoption markiert hat, oder dass der Interviewer eine solche Antwortvorgabe markiert hat.

Code	Label	Erläuterung zur Vergabe
-8	Weiß nicht/kann n. sagen	Missing aufgrund von Meinungslosigkeit oder Unwissenheit. Befragter kann nicht antworten, weil er die Antwort nicht weiß, oder nicht aus dem Gedächtnis abrufen kann
-9	Keine Angabe	Befragter hat eine Frage/ein Item gesehen, aber nicht beantwortet. Die Gründe hierfür sind häufig unbekannt (blank im Selbstausfüllerformat). Auch: Interviewer hat k.A. markiert, ohne dass Näheres zum Grund für fehlende Antwort bekannt ist (bspw. falls es für Interviewer keine Differenzierung zwischen "weiß nicht" oder "keine Angabe" gab)

2.2 Kategorien, die aus inhaltlichen Erwägungen als fehlend deklariert werden – „Inhaltlicher Trifft-nicht-zu-Code“

Auch bei dieser Kategorie haben Befragte Fragen zur Beantwortung gestellt bekommen, aber sich keiner Antwortkategorie zuordnen können, die als ‚gültiger Wert‘ vorgesehen war. Im Sinne einer sparsamen, das heißt Befragungszeit minimierenden, Fragebogenkonstruktion werden solche Befragte häufig ausgefiltert, bevor sie zu der entsprechenden Frage gelangen. Dies ist nicht immer möglich oder gewünscht, daher werden häufig bei den selbst oder durch den Interviewer zu selektierenden Ausfallgründen genauere Angaben erbeten, warum keine gültige Antwortkategorie selektiert werden kann. Genau diese Angaben sind es, die anschließend aus inhaltlichen Erwägungen als fehlend deklariert werden und denen hier entsprechende Missing-Codes zugewiesen werden sollen. Bei vielen Fragen kann es vorkommen, dass es nicht nur einen „inhaltlichen Trifft-nicht-zu-Code“ gibt, sondern mehrere solcher Codes. Um Informationsverlust zu minimieren ist es hier wünschenswert, die verschiedenen Ausfallgründe in die entsprechende Variable aufzunehmen, so dass für alle Variablen ein bestimmter Wertebereich für solche Ausfallgründe reserviert werden sollte. Auch einheitliche Variablenlabels sind natürlich nicht möglich, so dass diese entsprechend aus dem Erhebungsinstrument übernommen werden sollten.

Code	Label	Erläuterung zur Vergabe
-50 bis -59	Wird aus Fragebogen entnommen	In der Frage werden Informationen erfragt, die den Befragten nicht betreffen (inhaltlich: trifft nicht zu), aber Befragter wurde nicht ausgefiltert. Die entsprechenden Antwortlabel werden aus dem Fragebogen entnommen

Zwei Beispiele sollen die Vergabe der „inhaltlichen Trifft-nicht-zu-Codes“ illustrieren:

Beispiel 1:

*Frage*text: „Als Sie 15 Jahre alt waren, welche berufliche Stellung traf damals auf Ihren Vater zu? Ordnen Sie es bitte nach dieser Liste ein“

Gültiger Wertebereich: 10 Selbständiger Landwirt mit Fläche bis unter 10 ha
[...]
74 Praktikant/Volontär

Fehlende Werte: -55 Vater damals Rentner
-54 Vater damals arbeitslos
-53 Vater damals im Krieg
-52 Vater lebte nicht mehr
-51 Vater aus anderen Gründen n. erwerbstätig
-50 Vater unbekannt
-8 Weiß nicht/kann n. sagen
-9 Keine Angabe

Beispiel 2:

*Frage*text: „Wie hoch ist Ihr EIGENES monatliches Netto-Einkommen? Ich meine dabei die Summe, die nach Abzug der Steuern und Sozialversicherungsbeiträge übrig bleibt.“

Gültiger Wertebereich: 1 bis genanntes Maximum

Fehlende Werte: -50 Habe kein eigenes Einkommen
-7 Explizite Verweigerung
-9 Keine Angabe

Eine Kodierung auf einen ‚allgemeineren Ausfallgrund‘ wäre prinzipiell möglich, indem die verschiedenen Gründe, weshalb keine gültige Antwortkategorie zutraf, auf einen allgemeinen „Trifft nicht zu“-Code rekodiert werden (-10 ‚trifft nicht zu‘, vgl. folgendes Kapitel), dies dürfte aber nicht häufig gewünscht sein. Zudem erlaubt eine solche Rekodierung dann keinen Rückschluss allein anhand des Missing-Codes mehr darüber, welche Fragen einem Befragten zur Beantwortung präsentiert wurden, und welche nicht.

2.3 Fehlende Angaben, weil einem Teil der Befragten die entsprechende Frage nicht gestellt wurde („Filter-Missings“)

Alle unter diese Oberkategorie fallenden Codes haben die Gemeinsamkeit, dass Befragte keinen gültigen Wert bei einer Variablen haben, weil sie einer Teilpopulation angehören, der die entsprechende Frage im Fragebogen nicht gestellt worden ist. Grund hierfür kann eine Filterführung im Fragebogen sein oder die Nicht-Teilnahme an einer Teilstudie (z.B. ISSP). Streng genommen kann man auch die Nicht-Teilnahme an einer Teilstudie ebenfalls als Nichtteilnahme aufgrund einer Filterführung im Fragebogen verstehen, oder einen methodischen Split als einen speziellen Grund für eine Filterführung ansehen. Tiefergehende Informationen zu möglichen Befragteninformationen in anderen Split-Varianten oder über die Teilnahme am ISSP-Modul erscheinen jedoch wünschenswert, so dass für den ALLBUS die folgenden Missing-Kategorien gewählt werden.

Code	Label	Erläuterung zur Vergabe
-10	Trifft nicht zu	Befragter hat (aufgrund seiner vorherigen Angaben) eine Frage im Fragebogen nicht gestellt bekommen.
-11	TNZ (Fragebogen-Split)	Befragter hat aufgrund eines methodischen oder inhaltlichen Splits eine Frage oder Itematterie nicht gestellt bekommen
-12	TNZ (ISSP-Split)	Befragter befindet sich im jeweils anderen ISSP-Split
-13	keine ISSP-Teilnahme	Befragter hat am ISSP nicht teilgenommen
-14	TNZ (Zufallsauswahl)	Item wurde während des Interviews per Zufallsauswahl, d.h. ohne Dokumentation in einer Splitvariable, für den Befragten ausgefiltert

Bei den „Filter-Missings“ ist weiter zu beachten, dass Befragte – sollten mehrere Filterungen in einander verschachtelt sein – bei Folgefragen auf den Codes bleiben, denen sie in Vorfragen bereits zugeordnet sind. Das heißt, dass es bei einer Variable mehrere Filter-Missing Codes geben kann. Dies ist zum Beispiel dann der Fall, wenn innerhalb eines methodischen Splits verschiedenen Teilstichproben jeweils unterschiedliche Fragen gestellt werden. Dann erhalten diejenigen Personen, die sich nicht in dieser methodischen Splitversion befinden, den Code -11, diejenigen, die im Split aus inhaltlichen Gründen ausgefiltert werden, den Code -10. Ein weiteres Beispiel ist die Filterung im ISSP-Modul: Befinden sich Personen nicht im entsprechenden ISSP-Modul (sondern im anderen ISSP-Modul), bekommen sie den Code -12, Personen, die aus inhaltlichen Gründen in diesem ISSP-Modul ausgefiltert wurden, erhalten den Wert -10. Veranschaulicht könnte man auch sagen, dass „Filter-Missings“ die einzigen Codes für fehlende Werte sind, die über Folgefragen *und* generierte Variablen hinweg ‚durchgezogen‘ werden, also die Teilnehmer auf den Codes bleiben, auf denen sie in Vorfragen bereits kodiert waren.

Ein Beispiel soll die Anwendung dieses Missing-Schemas verdeutlichen:

Beispiel 3:

*Frage*text: „Sind/Waren Sie für die Arbeit anderer Mitarbeiter verantwortlich?“

Gültiger Wertebereich: 1 Ja
2 Nein

Fehlende Werte:	-13 keine ISSP-Teilnahme
	-12 TNZ (ISSP Split)
	-10 TNZ (ausgefiltert)
	-9 Keine Angabe

2.4 Missings nach Datenprüfung und -bereinigung

Die ALLBUS-Daten werden nach Abschluss der Erhebung ausführlichen Datenkontrollen und -prüfungen unterzogen. Gelegentlich kann es hierbei vorkommen, dass Fälle auf einen fehlenden Wert kodiert werden müssen. Hierfür wird der Code -41 „Datenfehler“ verwendet. Dies kann ein logisch unmöglicher oder empirisch hochgradig unplausibler Wert sein (z.B. eine tägliche Fernsehdauer von über 24 Stunden oder eine Aufgabe des letzten Hauptberufs mit 12 Jahren), oder ein Wert außerhalb des gültigen Wertebereichs einer Variable (z.B. Wert 5 wenn Variable einen Range von 1 bis 4 hat). Grundsätzlich ist festzustellen, dass solche Rekodierungen bei der Datenbereinigung äußerst sparsam vorgenommen werden, da seitens des Datenservice nachträgliche Plausibilitätskontrollen nur in geringfügigem Maße vorgenommen werden und nur dann korrigiert werden, wenn die Angaben offensichtlich falsch sind. Ansonsten bleibt es den Nutzern überlassen, entsprechende Plausibilitätskontrollen selbst vorzunehmen (etwa auch im Hinblick auf konsistente Angaben bei vorherigen oder späteren Fragen). Auch andere Gründe können für die Löschung eines Werts im Zuge der Datenaufbereitung sprechen (z.B. eine Angabe des Interviewers, dass es sich bei einem Wert um eine Fehleingabe handelt). Zudem kann im Zuge der Datenprüfung auffallen, dass Befragte Fragen erhalten haben, die für diese Befragte ausgefiltert hätten werden müssen (z.B. Fragen zur Erwerbstätigkeit bei Erwerbslosen), oder umgekehrt, dass Befragte Fragen nicht erhalten haben, die ihnen eigentlich hätten gestellt werden sollen. Auch hier werden die entsprechenden Fälle mit -41 „Datenfehler“ gekennzeichnet.

Code	Label	Erläuterung zur Vergabe
-41	Datenfehler	Ein Fehler bei der Datenerhebung oder -verarbeitung macht die Löschung der entsprechenden Werte notwendig (z.B. logisch unmöglicher oder empirisch hochgradig unplausibler Wert oder Kombination von Werten), bzw. Informationen liegen aufgrund eines Datenfehlers oder einer falschen Filterführung nicht vor.

2.5 Technisches Missing

Zu technischen Missings zählen Datenverluste aufgrund von ‚Technikversagen‘. Dies können Datenverluste auf bestimmten Befragungslaptops sein, eine falsch abgespeicherte Befragungsurzeit oder – falls solche Messungen vorgenommen wurden – fehlende Messungen der Antwortgeschwindigkeit. Hiervon zu unterscheiden sind Angaben des Interviewers, die wie Befragtenmerkmale behandelt werden sollen (wenn der Interviewer einen Kurzfragebogen zum Interview ausgefüllt hat oder Angaben wie die Befragungsdauer *notiert* bzw. in den Fragebogen *eingetragen* hat).

Code	Label	Erläuterung zur Vergabe
-88	Angabe fehlt	Datenverlust bei CAPI/CASI

2.6 Missings bei generierten und abgeleiteten Variablen

Im Folgenden soll die Missing-Vergabe bei generierten und abgeleiteten Variablen genauer erläutert und anhand von konkreten Beispielen beschrieben werden. Generierte und abgeleitete Variablen sind solche Variablen, die keine unmittelbare Entsprechung im Fragebogen haben, sondern nach der Befragung durch den ALLBUS-Nutzerservice oder das Befragungsinstitut gebildet werden. Generierte und abgeleitete Variablen erhalten zum Teil eigene Missing-Codes. Drei Arten von generierten Variablen sollen unterschieden werden:

1. Kategorisierungen und gruppierte Variablen
2. Kodierungen aus offenen oder halboffenen Fragen
3. Komplexere Variablenkonstruktionen

2.6.1 Missings bei Kategorisierungen und gruppierten Variablen

Kategorisierte und gruppierte Variablen erhalten keine eigenen Missing-Codes, sondern sollen abgesehen von der Gruppierung weiterhin den (zumeist kontinuierlichen) Ursprungsvariablen unmittelbar entsprechen. Dies soll den Nutzern einen schnellen Überblick über die Verteilung des interessierenden Merkmals, beispielsweise im Variable Report oder in ZACAT, bieten. Um Nutzern zu ermöglichen, gruppierte Variablen schnell und unkompliziert von Analysen und Variablenkonstruktionen auszuschließen (z.B. wenn eine Zählvariable über alle beantworteten Fragen generiert werden soll), werden diese Variablen im Zuge eines Konzepts für konstante Variablennamen im ALLBUS im Namen besonders gekennzeichnet.

2.6.2 Missings bei Kodierungen aus offenen oder halboffenen Fragen

Variablen, die (teilweise) aus den Antworten auf offene oder halboffene Fragen nachträglich kodiert werden, erhalten die aus dem Frage-Antwort-Prozess resultierenden Missing-Codes und zusätzlich einen eigenen Missing-Code.

- Da die Kodierung der Antwort aus einer (halb-)offenen Frage eine deutliche Entsprechung im Fragebogen hat, sollen die aus dem Frage-Antwort-Prozess resultierenden Missings der Ur-

ursprungsfrage im Fragebogen übernommen werden (z.B. „Keine Angabe“, „Weiß nicht“). Entsprechendes gilt auch, wenn bei einer Mehrfachnennung überhaupt keine Nennung gegeben wurde.

- Gab es bei der (halb-)offenen Frage eine Antwortmöglichkeit, die als „inhaltlicher Trifft-nicht-zu-Code“ behandelt wird, so wird dieser auch bei der kodierten Variable übernommen.
- Filter-Missings [-10, -14] werden aus der (halb-)offenen Ursprungsfrage übernommen. Werden Mehrfachnennungen kodiert, wird -10 „trifft nicht zu“ auch vergeben, wenn es keine weitere Nennung gibt.
- Als zusätzlicher Missing-Code wird bei kodierten Variablen der Code -33 „nicht bestimmbar“ vergeben, wenn die Antworten auf offene Nennungen unklar oder nicht lesbar sind, also keinem inhaltlichen Code zuordenbar sind.
- Sollte es im Einzelfall für eine bestimmte, eindeutige und lesbare Nennung keine Entsprechung im Codeschema geben, wird im Rahmen des ALLBUS stets versucht, dennoch einen inhaltlichen sinnvollen Code zu vergeben (ZUMA-Codes zur Ergänzung des ISCO, oder auch eine „Sonstiges“-Kategorie, z.B. bei Variable zu chronischen Krankheiten). Sollte dies im Einzelfall nicht möglich sein, sollte es also eine lesbare Nennung geben und für dieses Merkmal keinen entsprechenden Code auf der Zielvariablen, soll der Code -32 „nicht generierbar“ vergeben werden (siehe hierzu nächster Abschnitt).

Code	Label	Erläuterung zur Vergabe
-6 bis -9	Label des jeweiligen Missing-Codes	Wird aus Ursprungsfrage übernommen
-50 bis -59	Label des jeweiligen Missing-Codes	Wird aus Ursprungsfrage übernommen
-10 bis -14	Label des jeweiligen Missing-Codes	Wird aus Ursprungsfrage übernommen Bei Mehrfachnennungen -10 „trifft nicht zu“ zusätzlich auch für den Fall „keine weitere Nennung“
-33	Nicht bestimmbar	Wenn Antworten auf offene Nennungen unklar oder nicht lesbar sind

Beispiel 4:

Generierte Variable: Klassifikation des derzeitigen Berufs nach ISCO-88

Gültiger Wertebereich:	1 Soldat	
	[...]	
	9333 Frachtarbeiter	
Fehlende Werte:	-33 nicht bestimmbar	←
	-10 Trifft nicht zu	←

Befragter hat etwas angegeben, aber aus den (unvollständigen, ungenauen) Informationen ist kein ISCO-Code ableitbar

Befragter ist nicht berufstätig, Fragen zur Berufstätigkeit wurden entsprechend nicht gestellt

-9 Keine Angabe



Keine Angabe auf offene Ursprungsfragen zu beruflicher Tätigkeit und Namen des Berufs

-32 wird hier nicht vergeben, da für Berufe, die in die ISCO-Klassifikation nicht einordnenbar wären (z.B. „Soldat“), eigene ZUMA-Codes entwickelt wurden.

Beispiel 5:

Generierte Variable: Staatsangehörigkeit des Lebenspartners, erste Nennung

Gültiger Wertebereich:	0	Deutschland	
	[...]		
	545	Mikronesien	Entsprechender Ausfallgrund im Frageprogramm angegeben
Fehlende Werte:	- 50	Staatenlos	↙ Antwort nicht kodierbar, z.B. „Südamerika“, „Iraj“
	-33	nicht bestimmbar	←
	-10	Trifft nicht zu	← Kein Lebenspartner
	- 9	Keine Angabe	← Wenn keine Angabe Wenn „anderes Land“ genannt, aber keine offene Nennung

2.6.3 Missings bei komplexeren Variablenkonstruktionen

Variablen, die aufgrund einer Rechenoperation gebildet werden, und/oder keine simple Entsprechung im Fragebogen haben, werden hier als „komplexe“ generierte Variablen bezeichnet. Bei komplexen generierten Variablen werden keine Missing-Codes aus dem Frage-Antwort-Prozess vergeben. Hat der Befragte *auf mindestens einer* Ursprungsvariablen ein Filter-Missing [-10, -14], so wird bei der Ziel-Variablen der entsprechende Missing-Code vergeben. Dies soll Nutzern genauere Informationen über den Ausfallgrund geben, beziehungsweise Aufschluss darüber, ob ggf. für eine Variablenkonstruktion benötigte Merkmale in anderen Variablen abgelegt sind. Hat ein Befragter *unterschiedliche* Filter-Missings bei den Ursprungsvariablen oder gibt es bei einer Ursprungsvariablen einen „inhaltlichen Trifft-nicht-zu-Code“ [-50, -59] und bei anderen Ursprungsvariablen Filter-Missings [-10, -14], so wird als allgemeiner Ausfallgrund das Filter-Missing (zumeist -10 „Trifft nicht zu“) vergeben. Als eigener Missing-Code wird bei komplexeren generierten Variablen -32 „nicht generierbar“ vergeben. Die folgenden Missing-Codes werden bei generierten Variablen entsprechend vergeben:

Code	Label	Erläuterung zur Vergabe
-10 bis -14,	Label des jeweiligen TNZ-Codes	Werden aus den an der Variablenkonstruktion beteiligten Ursprungsvariablen übernommen. Bei verschiedenen TNZ-Codes auf Ursprungsvariablen wird das Filter-Missing (zumeist -10 TNZ) als allgemeinsten Ausfallgrund vergeben
-50 bis -59		

Code	Label	Erläuterung zur Vergabe
-32	Nicht generierbar	<p>Es fehlen auf einem oder mehreren Merkmalen Informationen, um die generierte Variable zu erstellen. Hierzu zählen auch Antwortverweigerungen oder nicht lesbare offene Nennungen bei den Ursprungsvariablen.</p> <p>Dieser Code wird ebenfalls vergeben, wenn die zu generierende Variable für ein Merkmal oder eine Merkmalskombination nicht definiert ist, gültige Nennungen auf der Ursprungsvariablen also nicht in einen Code der Zielvariablen überführbar sind.</p>

Für diese Variablengruppe sollen zwei Beispiele die Vergabe illustrieren.

Beispiel 6:

Generierte Variable: Alter der zweiten Person im Haushalt

Gültiger Wertebereich: 0 Unter 1 Jahr

[...]

95

Fehlende Werte:

-32 Nicht generierbar

-10 Trifft nicht zu



Keine Altersinformationen zur zweiten Haushaltsperson

Keine zweite Haushaltsperson vorhanden

Beispiel 7:

Generierte Variable: Body-Mass-Index

Gültiger Wertebereich: 13

[...]

63

Fehlende Werte:

-32 Nicht generierbar

-12 TNZ (ISSP-Split)

-13 keine ISSP-Teilnahme



Keine Antwort auf Frage nach Gewicht und/oder Körpergröße

Fragen nach Größe und/oder Gewicht nur in dem ISSP-Split, an dem Befragter nicht teilgenommen hat

Fragen nach Größe und/oder Gewicht nur im ISSP-Split; Befragter hat am ISSP nicht teilgenommen

2.7 Nicht enthaltene Informationen

2.7.1 Informationen, die bewusst nicht bereitgestellt werden

Vereinzelt werden Informationen nicht in den Datensätzen im „regulären“ Datenvertrieb bereitgestellt. Zumeist betrifft dies Informationen, die aus Datenschutzgründen zurückgehalten werden. In der ALL-BUS-Kumulation kann dies auch Variablen betreffen, die für neuere Querschnitte nicht mehr generiert

werden. Werden Informationen von Befragten oder Variablen in einem Datensatz bewusst nicht bereitgestellt, wird der Ausfall auf den Code -34 „nicht enthalten“ codiert.

Code	Label	Erläuterung zur Vergabe
-34	Nicht enthalten	Es gibt die Information anderswo oder die Variable wäre prinzipiell generierbar, wird aber in diesem Datensatz für Nutzer nicht zur Verfügung gestellt

2.7.2 Zusätzliches Missing für ALLBUS-Kumulation

In der ALLBUS-Kumulation sind alle Variablen enthalten, die mindestens zu zwei Erhebungszeitpunkten des ALLBUS erhoben worden sind. Aber die wenigsten Merkmale liegen für ausnahmslos alle ALLBUS-Erhebungen vor. Entsprechend wird ein Missing-Code für die ALLBUS-Kumulation benötigt, aus dem hervorgeht, dass einem Befragten bei der entsprechenden Variablen kein gültiger Wert zugeordnet werden kann, weil der Befragte Teilnehmer einer ALLBUS-Erhebung war, in der dieses Merkmal nicht erhoben wurde.

Code	Label	Erläuterung zur Vergabe
-1	Nicht erhoben	Nur in Kumulation: Variable ist bei Befragten nicht erhoben worden

So wie auch bisher, soll dieser Missing-Code mit „nicht erhoben“ belabelt werden. Als Code wird -1 festgelegt, damit dieser Code im Variable Report aufgrund der aufsteigenden Sortierung der Codes besonders ins Auge fällt.