Hinweise zur Codierung fehlender Werte in der Aufbereitung quantitativer Daten

fdb*info* Nr. 6 // Februar 2019 // Version 1.0

Text erstellt von **Karoline Harzenetter** (GESIS) in Zusammenarbeit mit **Claudia Neuendorf** (IQB), **Lisa Pegelow** (IQB) und **Ute Hoffstätter** (DZHW)

Bitte zitieren als: Verbund Forschungsdaten Bildung (2019): Hinweise zur Codierung fehlender Werte in der Aufbereitung quantitativer Daten. Version 1.0, fdb*info* Nr. 6.

1. Vorbemerkungen

Grundsätzlich sollten bei der Aufbereitung quantitativer Daten fehlende Werte in einem Datensatz mit der gleichen Sorgfalt behandelt werden wie dessen gültige Werte, d. h., fehlende Werte sollten ebenfalls codiert, benannt, auf Inkonsistenzen überprüft und als solche deklariert werden (vgl. Trixa et al. 2019; Jensen et al. im Druck; Jensen 2012: 31 f.). Die Dokumentation fehlender Werte in einem Datensatz erhöht nicht nur dessen Analysepotenzial, sondern auch die Qualität von Analyseergebnissen. Nur auf Basis ausführlicher Dokumentation der Ursachen fehlender Daten können Forschende überhaupt erst über die Zufälligkeit von Ausfällen entscheiden und den richtigen Umgang, z. B. in Form von Gewichtung oder Imputation (siehe auch Spieß 2010), mit diesen Werten wählen. Mögliche Ursache fehlender Informationen kann z. B. die Verweigerung einer Antwort durch den Befragten sein, der Einsatz einer Filterfrage, die Abwesenheit der Zielperson oder ihre Verweigerung der Teilnahme oder eine durch den Interviewer verursachte fehlerhafte Codierung usw. Antwortausfälle werden in der Regel bereits in der Konzeptionierung des Fragebogens und bei dessen Einsatz mitgedacht und eingeplant, können aber auch erst während und nach der Feldphase auftreten und sollten spätestens im Datensatz sichtbar und nachvollziehbar gemacht werden. Als Hilfestellung zur Codierung hat der VerbundFDB diese Handreichung auf Basis angewandter Standards des Deutschen Zentrums für Hochschul- und Wissenschaftsforschung (DZHW), Forschungsdatenzentrums am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) und des GESIS Datenarchivs entwickelt. Ist die Übergabe von Forschungsdaten an eines dieser Datenzentren geplant, ist es empfehlenswert, die jeweiligen dort institutsintern etablierten Codierungsstandards zu berücksichtigen. Die Standards können selbstverständlich auch auf Daten angewendet werden, die nicht bei einem dieser Institute archiviert werden.

2. Grundlegende Empfehlungen für die Codierung fehlender Werte des VerbundFDB

Fehlende Werte sollten aufgefächert und transparent in den Forschungsdaten dargestellt und dokumentiert werden. Aufgrund unterschiedlicher Umfragemethoden (z. B. online, telefonische oder persönliche Befragung), der großen Bandbreite eingesetzter Mittel (wie Software, Hardware oder

Papierfragebogen) und vielfältigen Ursachen fehlender Antworten ist dabei der Einsatz unterschiedlicher Kategorien notwendig. Manchmal sind auch durch Restriktionen der Software nur bestimmte Wertebereiche für fehlende Daten nutzbar. Entsprechend können Codierungsschemata je nach Erhebungsverfahren und eingesetzter Mittel variieren und unterschiedlich umfangreich sein. Trotz dieser Unterschiede sollten sich alle Schemata zur Kodierung fehlender Werte durch eine klare Abgrenzung zwischen gültigen und fehlenden Werten auszeichnen. Diese Abgrenzung kann bereits einfach durch die Verwendung negativer Zahlenbereiche für die fehlenden Werte, im Gegensatz zum positiven Wertebereich gültiger Antworten, erreicht werden. Die Verwendung von Werten, die deutlich außerhalb des validen Bereichs liegen, kann ebenfalls zur sichtbaren Abgrenzung dienen. Wenn beispielsweise einstellige Zahlen die Spannweite gültiger Antworten abdecken (Werte von 1-9), sollten Codes für fehlende Werte gewählt werden, die im oberen zweistelligen Bereich (z. B. 97, 98, 99) liegen. Der Wertebereich sollte so gewählt werden, dass er für alle Variablen eines Datensatzes gleichermaßen anwendbar ist. Beispielsweise steht dann der Wert „97“ bei allen Variablen eines Datensatzes für die Antwort „keine Angabe“ und variiert nicht von Variable zu Variable in einer Datendatei. Eine fein abgestufte Darstellung auftretender Arten fehlender Werte sollte außerdem gewährleistet werden.

Die Dokumentation fehlender Werte ist von immenser Bedeutung für die Datenqualität. Idealerweise sind die Kategorien im Datensatz mit sprechenden Labels versehen und ihre Bedeutung in Skalenhandbuch oder Methodenbericht umfassend erläutert. Häufig stellen Datengebende auch sogenannte Flagvariablen in ihren Datensätzen zur Verfügung. Diese Variablen markieren („flaggen“) Personen, deren Daten beispielsweise unvollständig sind oder die an einer kompletten Erhebung nicht teilgenommen haben. Diese Variablen ermöglichen es auch, verschiedene Gründe für den Ausfall anzugeben, ohne jeweils eine Vielzahl eigener, neuer Missing-Kategorien erstellen zu müssen. Sie erleichtern es Nutzenden eines Datensatzes, schnell einen variablenübergreifenden Überblick über die Struktur fehlender Daten zu gewinnen und diese als Filtervariablen zu verwenden, um ggf. bestimmte Fälle über die Flagvariable auszuschließen.

3. Codierungsstandards einzelner Forschungsdatenzentren (DZHW, IQB und GESIS)

Das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) nutzt für die Codierung fehlender Werte einen dreistelligen negativen Wertebereich (siehe Tabelle 1). In Anlehnung an das Nationale Bildungspanel (NEPS) lassen sich die Werte fünf verschiedenen Gruppen zuordnen. In den ersten beiden Gruppen wird zwischen fehlenden Werten aufgrund von Nicht-Beantwortung von Fragen seitens der Befragten (Nonresponse) und fehlenden Werten aufgrund der Filterführung bzw. für Befragte nicht relevanten Fragen unterschieden (Nicht zutreffend). Die dritte Gruppe beinhaltet Missingcodierungen, die im Zuge der Datenaufbereitung vergeben wurden (Editierter fehlender Wert). Die vierte Gruppe umfasst itemspezifische Missingcodierungen, die im Rahmen der Datenaufbereitung eines konkreten Datensatzes nur für einzelne Variablen vergeben wurden. Das DZHW unterscheidet dabei Missing-Kategorien, die nur bei Studien mit bestimmten Forschungsdesigns (z. B. Panel oder Querschnittsstudie) oder bei bestimmten Erhebungsmethoden anwendbar sind. So gibt es unterschiedliche mögliche Ursachen fehlender Werte bei selbst ausgefüllten schriftlichen (PAPI – paper and pencil interview) oder Online-Befragungen. Die fünfte Gruppe umfasst erhebungs- oder

datensatzspezifische fehlende Werte, die nicht durch die anderen Gruppen abgedeckt werden (andere fehlende Werte).

Das Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (FDZ am IQB) empfiehlt, einen Differenzierungsgrad in der Dokumentation fehlender Werte anzuwenden, der die fünf in Tabelle 2 dargestellten Kategorien fehlender Werte unterscheidet. Wichtiger als die genutzten Codes sind dabei eine einheitliche Anwendung innerhalb eines Datensatzes und über die Datensätze eines Projekts oder einer Studie hinweg und die Dokumentation in Form von sprechenden Labels im Datensatz und Erläuterungen im Skalenhandbuch sowie ggf. im technischen Bericht.

GESIS hingegen hat für die Codierung fehlender Werte keinen festgeschriebenen Standard, denn die Codierung von fehlenden Werten variiert zwischen den dort archivierten Studienreihen und Umfrageprogrammen oftmals, da diese meist ihren eigenen Standard entwickelt haben, wie beispielsweise die Studienreihe „Allgemeine Bevölkerungsumfrage Sozialwissenschaften (ALLBUS)“ (siehe Schulz o. J.). In dieser Handreichung sind von GESIS deshalb lediglich Codierungsbeispiele (Tabelle 3) aufgeführt, die in Studien Anwendung finden und an denen man sich bei der Aufbereitung und Konzeptionierung der eigenen Studie orientieren kann.

Tabelle 1: Empfehlung des FDZ des DZHW für die Codierung fehlender Werte:

Code	Label englisch/deutsch	Erläuterung
-999 bis -990: Item-Non-response		
-999	don't know / weiß nicht	
-998	no answer / keine Angabe	Befragte Person hat keine Angabe gemacht, also nichts angekreuzt (PAPI) bzw. keine Eingabe getätigt (Online).
-997	no answer (response category) / keine Angabe (Antwortkategorie)	Die befragte Person hat die Antwortkategorie „keine Angabe“ genutzt, die im Erhebungsinstrument (PAPI oder Online) explizit vorgesehen war.
-996[O]	interview break-off / Interviewabbruch	Expliziter Interviewabbruch
-995[P]	not participated (panel) / keine Teilnahme (Panel)	
-994	Refused / verweigert	nur wenn explizit eine Antwortoption "verweigert" ausgewählt werden kann.
-989 bis -970: nicht zutreffend oder nicht anwendbar		
-989	filtered / filterbedingt fehlend	Fehlender Wert aufgrund von Filterführung des Fragebogens.

Code	Label englisch/deutsch	Erläuterung
-988	does not apply / trifft nicht zu	Dieser Code wird nicht bei Filterführung vergeben, sondern wenn explizit eine Antwortoption "trifft nicht zu" vorgesehen ist.
-987	missing by design (questionnaire split) / designbedingt fehlend (Fragebogensplit)	
-986[P]	missing by design (wave) / designbedingt fehlend (Welle)	Frage wurde in dieser Welle nicht gestellt. Relevant für (Panel-)Daten im long-Format.
-985[Q]	missing by design (cohort) / designbedingt fehlend (Kohorte)	Frage wurde dieser Kohorte nicht gestellt. Relevant für gepoolte Datensätze.
-969 bis -950: editierter fehlender Wert (zu fehlend recodiert)		
-969	unknown missing / unbekannter fehlender Wert	Restkategorien, wenn kein anderes Missing rekonstruiert werden kann.
-968	implausible value / unplausibler Wert	Sollte möglichst nur verwendet werden, wenn bereits durch das Primärforschungsprojekt vergeben. Sollte nur sehr sparsam verwendet werden. Diese Entscheidung sollte möglichst den Datennnutzer(inne)n überlassen werden.
-967	anonymized / anonymisiert	Wird vergeben, wenn die Variable in dem jeweiligen Zugangsweg nicht verfügbar ist.
-966	not determinable / nicht bestimmbar	z.B. offene Angabe, die nicht vercodet werden konnte z.B. zwischen zwei Kästchen angekreuzt z.B. auch für generierte Variablen, die zwar in der Originalvariable einen gültigen Wert haben, für die in der generierten Variable jedoch keine Zuordnung stattfinden kann
-965	invalid multiple answer / ungültige Mehrfachnennung	Wenn Mehrfachantwort technisch nicht verhindert wurde.
-949 bis -930: Item-spezifische fehlende Werte mit informativen Wertelabels		
-929 bis -920: andere fehlende Werte		
-929	loss of data / Datenverlust	z.B. Werte aufgrund technischer Probleme nicht erfasst oder nicht erfragt.

Anmerkungen:
O = nur Onlinebefragungen
P = nur Panelbefragungen
Q = nur Querschnittsbefragungen

Tabelle 2: Empfehlung des FDZ am IQB für die Codierung fehlender Werte

Code	Label	Erläuterung
-99	Missing Omitted / Missing by Intention	Item wurde nicht bearbeitet
-98	Missing Invalid Response	ungültige Antwort / nicht interpretierbar
-97	Missing By Design	Item nicht administriert/nicht bearbeitbar
-96	Missing Not Reached	Befragung wurde vor dem Item abgebrochen
-95	sonstige Missings	Grund für Missing unklar, Sysmis, Missingart konnte auch nach Rücksprache mit Datengebern nicht geklärt werden, anonymisierte Missings etc.

Tabelle 3: Codierungsbeispiele aus dem GESIS-Datenarchiv (DAS) (Jensen 2018: 22)

	Code	Label englisch	Label deutsch
Beispiel 1 Wenn „7“ ein gültiger Wert ist, wird „97“ codiert	7 (bzw. 97, 997)	refused	verweigert
	8 (bzw. 98, 998)	don‘t know	weiß nicht
	9 (bzw. 99, 999)	no answer	keine Angabe
	0	does not apply	trifft nicht zu
Beispiel 2 Wenn „0“ in den Bereich gültiger Werte fällt, wird „9“ als „trifft nicht zu“ codiert	6 (bzw. 96, 996)	refused	verweigert

	7 (bzw. 97, 997)	don‘t know	weiß nicht
	8 (bzw. 98, 998)	no answer	keine Angabe
	9 (bzw. 99, 999)	does not apply	trifft nicht zu
Beispiel 3 Codierung mit negativen Werten	-1	don‘t know	weiß nicht
	-2	no answer	keine Angabe
	-3	does not apply	trifft nicht zu
	-4	not asked in survey	nicht abgefragt

4. Referenzen

Jensen, Uwe; Netscher, Sebastian; Weller, Katrin (im Druck): Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten. Opladen, Berlin und Toronto: Budrich.

Jensen, Uwe (2018): Data Processing. In: Netscher, Sebastian; Eder, Christina (Hrsg.): Data Processing and Documentation: Generating high quality research data in quantitative social science research. GESIS Papers 2018/22. URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59492-3.

Jensen, Uwe (2012): Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten. GESIS Technical Reports 2012/07. URL: www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf.

Schulz, Sonja (o. J.): Kodierung und Definition von fehlenden Werten im ALLBUS – ein vereinheitlichtes Missing-Schema. URL: https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/allbus/dokumente/Kodierung_fehlender_Werte.pdf.

Skopek, Jan; Pink, Sebastian und Bela, Daniel (2013): Starting Cohort 4: Grade 9 (SC4). SUF Version 1.1.0 Data Manual (Federal Ministry of Education and Research, Hrsg.) (Research Data). Bamberg: National Educational Panel Study.

Spieß, Martin (2010): Der Umgang mit fehlenden Werten. In: Wolf, Christof; Best, Henning (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften.

Trixa, Jessica; Ebel, Thomas und Harzenetter, Karoline (2019): Hinweise zur Aufbereitung quantitativer Daten. forschungsdaten bildung informiert, Nr. 4.