

Einkommenserfassung in der Kommunalen Bürgerumfrage

Waschipky, Martin

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:

Waschipky, M. (2022). Einkommenserfassung in der Kommunalen Bürgerumfrage. *Stadtforschung und Statistik : Zeitschrift des Verbandes Deutscher Städtestatistiker*, 35(2), 66-73. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-81796-4>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY-SA Lizenz (Namensnennung-Weitergabe unter gleichen Bedingungen) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier: <https://creativecommons.org/licenses/by-sa/4.0/deed.de>

Terms of use:

This document is made available under a CC BY-SA Licence (Attribution-ShareAlike). For more information see: <https://creativecommons.org/licenses/by-sa/4.0>


Martin Waschipky

Einkommenserfassung in der Kommunalen Bürgerumfrage

Die Einkommenserfassung stellt eine schwierige, aber bedeutsame Herausforderungen in sozialwissenschaftlichen Erhebungen dar. Der Beitrag analysiert die Effekte des Erhebungsinstruments auf das Antwortverhalten der Befragten. Es zeigt sich, dass die offene Abfrage zu deutlich höherer Antwortverweigerung führt. Darüber hinaus werden Hinweise und Empfehlungen zum Umgang mit den gewonnenen Daten gegeben.

Martin Waschipky, M.A.

Kommunikations- und Medienwissenschaft, seit 2019 wissenschaftlicher Sachbearbeiter in der Abteilung Stadtforschung im Amt für Statistik und Wahlen der Stadt Leipzig. Schwerpunkte sind unter anderem: Datenaufbereitung, -analyse und Visualisierung vor allem mit R, explorative und schließende Statistik und der Mietspiegel der Stadt Leipzig.

 martin.waschipky@leipzig.de

Schlüsselwörter

Nettoäquivalenzeinkommen – Einkommenserfassung – Methodenexperiment – Datenaufbereitung – kommunale Bürgerumfrage

Einleitung

In vielen Umfragen wird das Einkommen der Befragten erhoben. Sowohl das persönliche Einkommen als auch das Haushaltseinkommen ermöglichen in der Auswertung wichtige Entwicklungen aufzuzeigen und Quoten zu berechnen. Beispielsweise stellen die Einkommensentwicklung, Mietbelastungs- sowie Armutsgefährdungsquote wichtige Indikatoren dar und sind Entscheidungsgrundlagen für Stadtpolitik und verwaltung, weshalb die Einkommenserhebung in kommunale Umfragen meist erforderlich ist.

Das Einkommen kann auf unterschiedliche Weise erfasst werden: klassiert (geschlossen), mit vorgegebenen Einkommensspannen zur Auswahl, oder spitz (offen), indem die Befragte ihre individuelle Einkommenssumme eintragen. Die verschiedenen Operationalisierungen bringen jeweils Herausforderungen mit sich (kognitive Belastung, sensitive Fragen, soziale Erwünschtheit u. a.), die in der Fachliteratur umfassend diskutiert wurden und werden.

Die Kommunale Bürgerumfrage (KBU) wird in Leipzig seit 1991 jährlich erhoben. Seit 2008 findet zudem regelmäßig eine Erhebung mit einer deutlich größeren Stichprobe statt, um auf Ebene der 63 Ortsteile valide Ergebnisse zu erhalten. Seit 2011 wird die Ortsteilbefragung im zweijährigen Turnus erhoben. Bis 2006 wurden das persönliche und Haushaltseinkommen dabei offen abgefragt. Die itemspezifische Antwortverweigerung nahm jedoch im Laufe der Zeit deutlich zu und erreichte mit 27 Prozent im Jahr 2006 den Negativrekord. Entsprechend wurde 2007 ein methodischer Wechsel hin zu klassierten Abfragen vollzogen.

Durch diesen Wechsel konnte (zumindest bezogen auf die Einkommensmessung) der Trend immer höherer Antwortverweigerung gestoppt werden. Da das höhere Datenniveau der spitzen Werte jedoch vielfältigere Auswertungsmethoden erlaubt und darüber hinaus auch fragebogenökonomische Überlegungen eine offene Abfrage charmant erscheinen lassen¹, wurde im Rahmen der Leipziger Kommunalen Bürgerumfrage 2021 ein Quasi-Experiment mit den beiden Erhebungsformen durchgeführt, um die in der Vergangenheit beobachtete Tatsache des höheren Item-Non-Response systematisch untersuchen zu können.

Das in diesem Artikel beschriebene Methodenexperiment² stellt kein umfassendes Forschungsvorhaben dar und geht entsprechend mit gewissen Einschränkungen einher. Gleichwohl besitzen die Befunde Praxisrelevanz für Stadtforsche-

rinnen und -forscher, hinsichtlich der Frage nach Vor- und Nachteilen der jeweiligen Erhebungsform.

Zur Prüfung der Forschungsfrage, ob eine offene Abfrage mit einer erhöhten Antwortverweigerung einhergeht, wurden in zwei von vier Fragebogenvarianten der KBU 2021 die Einkommen klassiert, in den beiden anderen spitz, abgefragt.

Im Folgenden werden zunächst die Instrumente sowie der allgemeine Rücklauf kurz vorgestellt. Anschließend folgt eine itemspezifische Non-Response-Analyse. Die Herausforderungen bei der Datenaufbereitung und -auswertung bilden einen weiteren Aspekt, welcher die Empfehlungen aus Leipziger Perspektive ergänzt. Das Fazit fasst die Befunde zusammen.

Erhebungsinstrumente und Rücklauf

Zunächst stellt Abbildung 1 die beiden Erhebungsinstrumente der KBU 2021 gegenüber. Dabei wird auch deutlich, was unter Nettoeinkommen zu verstehen ist. Sowohl für das Haushalts- als auch das persönliche Nettoeinkommen waren die Frageformulierungen identisch. Die Einkommensklassen waren zuletzt 2019 angepasst worden, da in den vergangenen Jahren deutliche Einkommenszuwächse zu verzeichnen waren. Zum einen wurde eine Ausdifferenzierung im mittleren, zum anderen eine Erhöhung im oberen Einkommensbereich erforderlich.

Grundsätzlich werden in den vier Fragebogenversionen (A–D) thematisch unterschiedliche Fragestellungen verfolgt. Bei der Variante D handelt es sich um die Ortsteilbefragung, welche einen deutlich höheren Stichprobenumfang aufweist. Hinsichtlich des fragebogenspezifischen Rücklaufs zeigt Tabelle 1, dass alle Versionen (A–D) auf in etwa gleichem Niveau liegen. Auch hinsichtlich der Fragevariante zum Einkommen (geschlossen vs. offen) zeigen sich keine signifikanten Unterschiede im Rücklauf bzw. Unit-Non-Response.

Itemspezifische Non-Responseanalyse

Wie bereits in Tabelle 1 dargestellt, liegt kein fragebogenspezifischer Non-Response vor. Entsprechend stellt sich die Frage, ob hinsichtlich der beiden Einkommensabfragen itemspezifische Antwortausfälle oder Verweigerungen auftreten. Im Folgenden werden die plausibilisierten Daten genutzt³ und auf die Teilmenge, welche repräsentativ für die Stadt Leipzig ist, fokussiert.⁴

Abbildung 2 stellt dabei das Antwortverhalten sowohl differenziert nach Einkommensart als auch Fragevariante gegenüber. Dabei wird deutlich, dass die offene Abfrage zu einer signifikant höheren Verweigerung führt. Der Effekt ist mit einem Cramers V von .13-.14 zwar nach Konvention als eher klein zu bezeichnen (vgl. Sun et al. 2010), doch handelt

Abb. 1: Items zur Einkommenserhebung der Kommunalen Bürgerumfrage 2021 Leipzig im Vergleich

Wie hoch ist das durchschnittliche Haushalts-Nettoeinkommen im Monat insgesamt? Bitte zählen Sie die €-Beträge aller Haushaltsmitglieder nach Abzügen zusammen, z. B. Lohn oder Gehalt, Unternehmereinkommen, Rente, Pension, öffentliche Unterstützungen (z. B. Arbeitslosengeld I, Arbeitslosengeld II („Hartz IV“), Sozialgeld, Kosten der Unterkunft, Wohngeld, Kindergeld, Krankengeld, Elterngeld, BAFöG), Einkommen aus Vermietung und Verpachtung. Wenn Sie in einer Wohngemeinschaft leben, geben Sie bitte nur die Personen als gemeinsamen Haushalt an, die gemeinsam wirtschaften.

■ unter 500 €	■ 1000 bis unter 1100 €	■ 2300 bis unter 2600 €	■ 4400 bis unter 4800 €
■ 500 bis unter 600 €	■ 1100 bis unter 1300 €	■ 2600 bis unter 2900 €	■ 4800 bis unter 5200 €
■ 600 bis unter 700 €	■ 1300 bis unter 1500 €	■ 2900 bis unter 3200 €	■ 5200 bis unter 5600 €
■ 700 bis unter 800 €	■ 1500 bis unter 1700 €	■ 3200 bis unter 3600 €	■ 5600 bis unter 6000 €
■ 800 bis unter 900 €	■ 1700 bis unter 2000 €	■ 3600 bis unter 4000 €	■ 6000 bis unter 7000 €
■ 900 bis unter 1000 €	■ 2000 bis unter 2300 €	■ 4000 bis unter 4400 €	■ 7000 € und mehr

Nettoeinkommen aller Haushaltsmitglieder:
(nur volle Euro, ohne Nachkommastelle) [] € pro Monat

Wie hoch ist ihr durchschnittliches persönliches Nettoeinkommen im Monat?

■ unter 150 €	■ 600 bis unter 700 €	■ 1400 bis unter 1600 €	■ 2600 bis unter 3000 €
■ 150 bis unter 300 €	■ 700 bis unter 800 €	■ 1600 bis unter 1800 €	■ 3000 bis unter 3400 €
■ 300 bis unter 400 €	■ 800 bis unter 1000 €	■ 1800 bis unter 2000 €	■ 3400 bis unter 3800 €
■ 400 bis unter 500 €	■ 1000 bis unter 1200 €	■ 2000 bis unter 2300 €	■ 3800 bis unter 4200 €
■ 500 bis unter 600 €	■ 1200 bis unter 1400 €	■ 2300 bis unter 2600 €	■ 4200 € und mehr

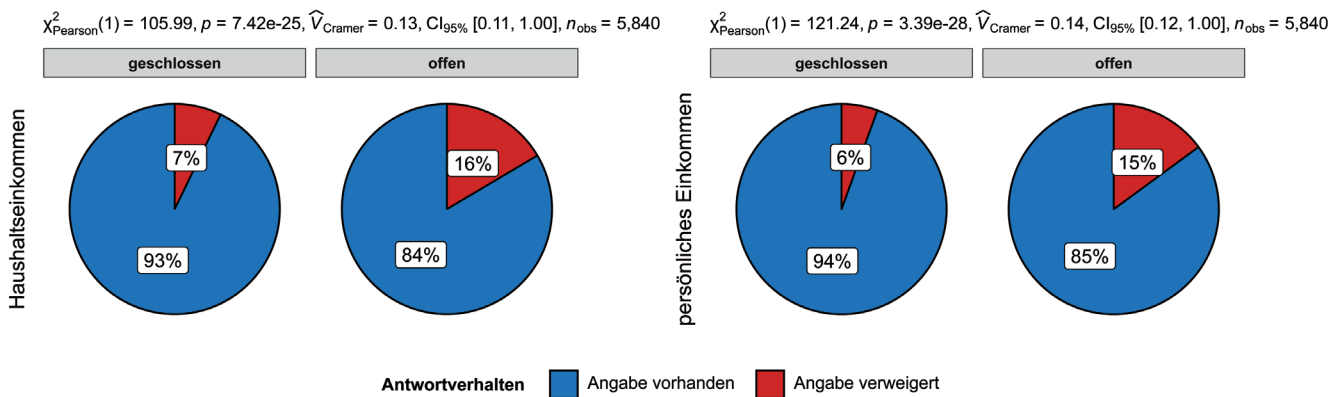
Ihr persönliches Nettoeinkommen (nur volle Euro, ohne Nachkommastelle) [] € pro Monat

Tab. 1: Rücklauf nach Fragebogen und Fragevariante absolut und relativ

Fragebogen (Fragevariante zum Einkommen)	Brutto- stichprobe	Neutrale Ausfälle		Nettostich- probe	auswertbare Datensätze			Rücklaufquote (Prozent)
		absolut	Prozent		Gesamt	darunter Online		
						absolut	Prozent	
A (geschlossen)	3.000	131	4,4	2.869	1.121	355	31,7	39,1
B (geschlossen)	3.000	138	4,6	2.862	1.117	341	30,5	39,0
C (offen)	3.000	125	4,2	2.875	1.153	327	28,4	40,1
D (offen)	16.000	358	2,2	15.642	6.217	1684	27,1	39,7
Gesamt	25.000	752	3,0	24.248	9.608	2.707	28,2	39,6

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Abb. 2: Itemspezifischer Non-Response nach Einkommensart und Fragevariante



Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

es sich um eine mehr als doppelte so hohe Verweigerung, was in der praktischen Anwendung als durchaus problematisch anzusehen ist, insbesondere dann, wenn diese Ausfälle mit weiteren Merkmalen in Beziehung stehen (z. B. Geschlecht, Bildung, berufliche Stellung etc.).

Tabelle 2 erlaubt eine tiefere Betrachtung dieser Zusammenhänge mittels schrittweiser Regressionsmodelle. Zunächst wurde aus allen Fragebogenvarianten eine Zufallsauswahl von jeweils 900 Datensätzen gezogen. Dieses Vorgehen war nötig, um reine Fallzahleneffekte aufgrund der disproportionalen Stichprobenziehung auszuschließen. Beispielsweise umfasst die Teilmenge Fragebogenvariante D doppelt so viele Fälle wie jeweils die der drei anderen Varianten (A–C). Hierdurch werden auch ähnlich starke Effekte unterschiedlich hinsichtlich ihres Signifikanzniveaus bewertet, was zu Fehlschlüssen führen kann. Weiterhin wurden Fälle ausgeschlossen, bei denen Missings in den im Modell enthaltenen Differenzierungsvariablen vorlagen.⁵ Um die Komplexität nicht weiter zu steigern, wurden zudem Fälle ausgeschlossen, welche nur eine der beiden Einkommensvariablen (persönliches oder Haushaltseinkommen) beantworteten (5 Prozent der Fälle).

Im Ergebnis stehen 3.600 Datensätze aus vier verschiedenen Fragebogenvarianten (A-D) zur Verfügung. In der schrittweisen logistischen Regression werden mit jedem Schritt weitere Variablen zur Erklärung der Verweigerung hinzugefügt.

Dabei geben die Odds Ratios (OR) die Chancenverhältnisse zu erkennen. Interpretieren lassen sich diese wie folgt: Ein OR größer 1 bedeutet eine Steigerung der Antwortwahrscheinlichkeit, ein OR kleiner 1 verringert diese bzw. steigert die Wahrscheinlichkeit der Verweigerung. Step 0 zeigt dabei den Rohwert an, welcher – umgerechnet in eine Wahrscheinlichkeit – bei 93 Prozent liegt. Durch die weiteren Schritte (Step 1 bis 4) wird deutlich, dass Zusammenhänge zwischen soziodemografischen Variablen und der Neigung zur Verweigerung bestehen. Unklar bleibt in dieser Analyse, ob die Verweigerungsneigung mit der Fragevariante zusammenhängt, oder unabhängig von dieser ist. Weiterhin ist die Güte des Modells sehr gering (Pseudo R² bei 2 Prozent). Überdies handelt es sich zum Teil um hochkorrelierte Variablen (Alter und Stellung im Erwerbsleben). Entsprechend dient diese Analyse lediglich als exploratives Instrument zur Aufdeckung von möglichen Beziehungen zwischen Variablen, welche in einem weiteren Schritt genutzt werden sollen.

Tab. 2: Schrittweise logistische Regression zur Identifizierung relevanter Differenzierungsvariablen

Predictors	Step 0			Step 1			Step 2			Step 3			Step 4		
	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
(Intercept)	13.23	11.64 – 15.03	<0.001	14.11	11.70 – 17.20	<0.001	22.73	17.54 – 29.93	<0.001	27.22	19.18 – 39.67	<0.001	39.01	25.47 – 62.27	<0.001
Geschlecht: weiblich				0.89	0.69 – 1.15	0.370	0.88	0.68 – 1.14	0.335	0.90	0.69 – 1.17	0.433	0.94	0.72 – 1.22	0.634
schulischer Abschluss: Fachabitur							0.73	0.46 – 1.19	0.188	0.80	0.50 – 1.32	0.360	0.76	0.48 – 1.27	0.284
schulischer Abschluss: 10.Klasse							0.46	0.34 – 0.62	<0.001	0.53	0.38 – 0.72	<0.001	0.48	0.35 – 0.67	<0.001
schulischer Abschluss: 8./9.Klasse							0.27	0.18 – 0.40	<0.001	0.35	0.23 – 0.53	<0.001	0.37	0.24 – 0.58	<0.001
Alter: 35 bis 54										0.85	0.57 – 1.27	0.437	0.64	0.40 – 1.00	0.057
Alter: 55 bis 90										0.61	0.41 – 0.89	0.011	0.68	0.39 – 1.18	0.172
Stellung im Erwerbsleben: arbeitslos													0.59	0.30 – 1.30	0.155
Stellung im Erwerbsleben: Rentner													0.54	0.35 – 0.81	0.004
Stellung im Erwerbsleben: Stud./Schüler													0.30	0.16 – 0.57	<0.001
Observations	3600			3600			3600			3600			3600		
R ² Tjur	0.000			0.000			0.015			0.018			0.024		

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Die Analyse in Tabelle 3 nutzt jene Befunde und erweitert die Perspektive: Geprüft wird, ob die auf Gesamtebene gefundenen Prädiktoren ihre Signifikanz unabhängig von der Frageversion behalten und ob spezifische Unterschiede zwischen den beiden Varianten der Einkommensabfrage bestehen. Hierfür wird die abhängige Variable (Beantwortung sowohl persönliches als auch Haushaltseinkommen) gegen die dargestellten Prädiktoren regressiert.

Dabei zeigt sich zum einen, dass die auf Gesamtebene gefundenen Unterschiede bei der geschlossenen Abfrage teilweise ihre Signifikanz verlieren (Stellung im Erwerbsleben), zum anderen ist die Wahrscheinlichkeit einer Beantwortung bzw. Nicht-Verweigerung bei der geschlossenen Abfrage deutlich höher (OR von 159 bzw. 99 Prozent vs. OR von 23 bzw. 96 Prozent⁶). Die Fragevariante hat somit den stärksten Einfluss auf die itemspezifische Antwortbereitschaft bzw. Verweigerung.

Tab. 3: Einfluss der Differenzierungsvariablen in Abhängigkeit der Fragevariante auf die Antwortbereitschaft

Predictors	Offene & geschlossene Abfrage			offene Abfrage			geschlossene Abfrage		
	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
(Intercept)	97.76	59.46 – 167.30	<0.001	22.69	14.09 – 38.35	<0.001	159.44	56.92 – 627.28	<0.001
Geschlecht: weiblich	0.95	0.73 – 1.24	0.724	0.97	0.71 – 1.31	0.835	0.94	0.54 – 1.61	0.820
schulischer Abschluss: Fachabitur	0.75	0.47 – 1.26	0.266	0.74	0.43 – 1.34	0.302	0.63	0.25 – 1.81	0.351
schulischer Abschluss: 10.Klasse	0.47	0.34 – 0.65	<0.001	0.46	0.31 – 0.67	<0.001	0.47	0.23 – 0.92	0.030
schulischer Abschluss: 8./9.Klasse	0.35	0.22 – 0.54	<0.001	0.34	0.20 – 0.57	<0.001	0.35	0.14 – 0.86	0.021
Alter: 35 bis 54	0.61	0.38 – 0.97	0.042	0.73	0.43 – 1.22	0.242	0.30	0.08 – 0.88	0.049
Alter: 55 bis 90	0.67	0.38 – 1.16	0.151	0.81	0.43 – 1.51	0.508	0.31	0.07 – 1.06	0.080
Stellung im Erwerbsleben: arbeitslos	0.58	0.29 – 1.30	0.153	0.55	0.25 – 1.39	0.165	0.72	0.20 – 4.64	0.669
Stellung im Erwerbsleben: Rentner	0.53	0.34 – 0.81	0.004	0.49	0.29 – 0.79	0.004	0.71	0.30 – 1.57	0.408
Stellung im Erwerbsleben: Stud./Schüler	0.28	0.15 – 0.54	<0.001	0.23	0.12 – 0.45	<0.001	0.78	0.12 – 15.45	0.827
Fragevariante: offen	0.25	0.18 – 0.34	<0.001						
Observations	3600			1800			1800		
R ² Tjur	0.054			0.039			0.014		

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Womit die erhöhte Verweigerung zusammenhängt, kann an dieser Stelle nicht beantwortet werden und ist auch nicht Teil der Fragestellung. Entsprechende Gründe werden und wurden in der Fachliteratur diskutiert und reichen von der höheren kognitiven Anforderung bei offener Abfrage bis hin zur gesteigerten Sensitivität einer spitzen Abfrage. Für die Stadtforscher/-innen bleibt am Ende jedoch der Befund, dass die spitze Abfrage die Itemverweigerung deutlich erhöht.

Datenaufbereitung und -auswertung

Unterschätzt wurden überdies die spezifischen Herausforderungen im Datenhandling, welche sich durch eine Kombination beider Erhebungsmodi ergeben. Zunächst wurde der Prozess entsprechend der gewohnten Routinen aufgesetzt: Die offenen Angaben wurden in klassierte Ausprägungen überführt – what could possibly go wrong?

Dabei zeigte sich in den ersten Ergebnisauswertungen, dass die Befragten, welche eine offene Abfrage erhielten, einen signifikant höheren Median aufwiesen als jene, welchen die klassierte Abfrage vorgelegt wurde (vgl. Tabelle 4)

In der weiteren Analyse wurde jedoch deutlich, dass der aus den Klassen berechnete Median nicht dem tatsächlichen Median, welcher anhand der offen abgefragten Werte bestimmt wurde, entspricht. Somit führt eine einfache Umwandlung der offenen Angaben in klassierte Werte zu einem falschen Ergebnis – ein Fakt, der weniger offensichtlich ist, als es aufgrund der Ausführungen anmuten mag.

In der Folge standen zwei Optionen zur Verfügung: Die kompliziertere, aber korrekte wird entsprechend in der Auswertung der Kommunalen Bürgerumfrage 2021 angewandt: Zur Bestimmung des städtischen Medians wird das arithmetische Mittel der klassierten und offenen Mediane, entsprechend nach Anzahl der befragten Personen gewichtet, bestimmt. Abb. 3 gibt dies mittels Dumbbell-Plots zu erkennen.

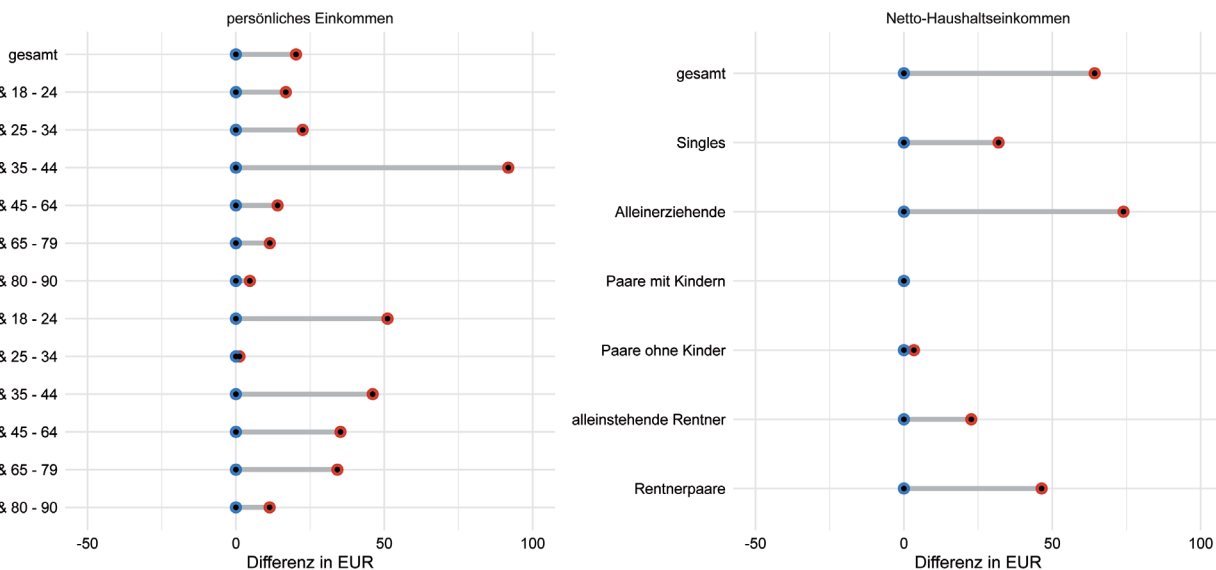
Tab. 4: Persönliches und Haushaltseinkommen nach Fragebogenversion und Fragevariante

Fragebogenversion (Fragevariante)	persönliches Nettoeinkommen		Netto-Haushaltseinkommen	
	Anzahl	Median in EUR	Anzahl	Median in EUR
A (geschlossen)	1.063	1.573	1.043	2.079
B (geschlossen)	1.051	1.588	1.034	2.076
C (offen)	997	1.649*	999	2.163*
D (offen)	2.067	1.626*	2.009	2.161*

*unkorrigierte klassierte Mediane

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Abb. 3: Differenzen zwischen unkorrigiertem und korrigiertem Median differenziert nach Geschlecht und Altersgruppe



Lesehinweis: Beim unkorrigierten Median wurden die offen abgefragten Einkommenswerte in Klassen überführt. Entsprechend wird ein Median für alle Fragevarianten auf Basis klassierter Werte berechnet. Der korrigierte Median stellt das gewichtete Mittel (anhand der Fälle) aus dem klassierten und dem Median der spitzen Werte dar und bildet den tatsächlichen Messwert ab.

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Dort sind die Differenzen zwischen dem Median, welcher durch die Überführung der offenen Angaben in klassierte Werte ermittelt wurde, und dem „korrigierten“ Median, welcher die klassierten und offenen Mediane mittels gewichtetem Mittelwert kombiniert, dargestellt. Die entsprechenden Ergebnisse zeigen die Werte auf Basis aller Fragebogenvarianten (A-D), auf Gesamtebene sowie für die Differenzierung nach Geschlecht und Altersgruppe.

Eine Alternativversion zur Zusammenführung offen und geschlossen abgefragter Einkommen, welche den Auswertungsprozess deutlich erleichtert, da die bisherigen Routinen – Auswertung auf Basis klassierter Werte – beibehalten werden können, sei an dieser Stelle als R-Praxis-Tipp angeboten:

```
#install.packages(„dplyr“)
library(dplyr)

#Klassengrenzen persönliches Einkommen
peki <- c (0,150,300,400,500,600,700,800,1000,1200,1400,
1600,1800,2000,2300,2600,3000,3400,3800,4200,20000)

#Klassengrenzen Haushaltseinkommen
heki <- c (0,500,600,700,800,900,1000,1100,1300,1500,
1700,2000,2300,2600,2900,3200,3600,4000,4400,4800,
5200,5600,6000,7000,20000)

set.seed (2022) #Reproduzierbarkeit
umf %>%

# um systematische Verzerrungen zu verhindern
dplyr::sample_n(nrow(.))%>%

# Datensatz durchnummerieren
dplyr::mutate(rowid = dplyr::row_number())%>%
dplyr::mutate(pek = dplyr::case_when(

# Klassierte Werte werden übernommen
!is.na(pek) ~ as.numeric(pek),

# abwechseln werden die Fälle in die Untergrenze
!is.na(pek_offen) & rowid %% 2 == 0 ~
as.numeric(cut(pek_offen, breaks = peki)),

# oder eine Klasse tiefer eingeordnet
!is.na(pek_offen) & rowid %% 2 == 1 ~
as.numeric(cut(pek_offen, breaks = peki + 1)))) %>%
mutate(hhek = case_when(

# Klassierte Werte werden übernommen
!is.na(hhek) ~ as.numeric(hhek),

# abwechseln werden die Fälle in die Untergrenze
!is.na(hhek_offen) & rowid %% 2 == 0 ~
as.numeric(cut(hhek_offen, breaks = heki)),

# oder eine Klasse tiefer eingeordnet
!is.na(hhek_offen) & rowid %% 2 == 1 ~
as.numeric(cut(hhek_offen, breaks = heki + 1))))
```

Im Fallbeispiel werden zunächst die Einkommen, welche oberhalb der höchsten Klassengrenze (20.000 Euro) liegen, auf diese begrenzt. Anschließend werden die Daten in eine zufällige Reihenfolge gebracht, um systematische Verzerrungen zu vermeiden. Jeder Datensatz wird fortlaufend nummeriert, wodurch über den Modulo Befragte, welche auf Klassengrenzen (1.000 Euro, 1.200 Euro etc.) liegen, abwechselnd nach oben und unten verschoben werden. Auf Befragte, welche nicht genau auf einer Klassengrenze liegen oder die ohne-

hin eine klassierte Abfrage erhielten, hat das Vorgehen keine Auswirkungen. Die Variablen heißen entsprechend hhek (Haushaltseinkommen) bzw. pek (persönliches Einkommen), der Zusatz „_offen“ identifiziert entsprechend jene Variablen, welche offene erhobene Werte enthalten.

Hierdurch nähern sich die Mediane deutlich dem Bereich des „wahren“ Medians an.⁷ Beim Haushaltseinkommen existiert insbesondere im Bereich um die 2.000 Euro eine Häufung von Fällen, welche sich offen abgefragt genau bei 2.000 Euro einordnen, geschlossen abgefragt sich jedoch bei 1.700 bis unter 2.000 Euro bzw. bei 2.000 bis unter 2.300 Euro einordnen würden. Beim persönlichen Einkommen zeigt sich dieser Effekt im Bereich um 1.400 Euro. Abbildung 4 macht dies mittels Histogrammen noch einmal deutlich: Die offenen Angaben, welche ohne Korrektur in Klassen überführt wurden, führen zu einer deutlichen Überschätzung des Medians. Demgegenüber stellt die einfache Zufallsverteilung der Befragten, welche genau auf einer Klassengrenze liegen, eine deutliche Verbesserung dar.

Einordnung und Empfehlung aus Perspektive der Leipziger Erhebung

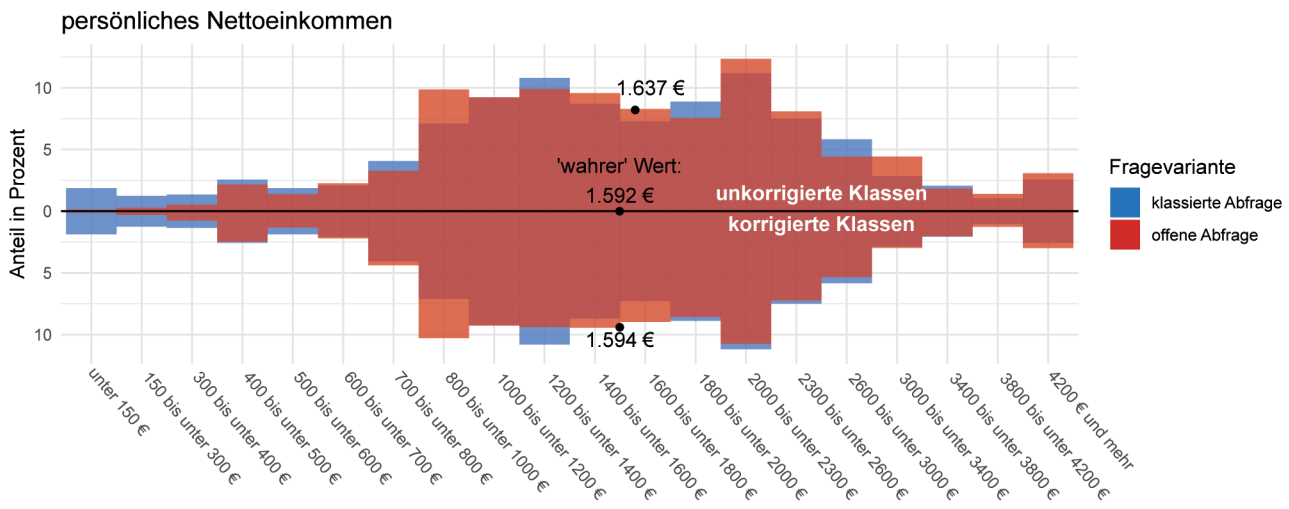
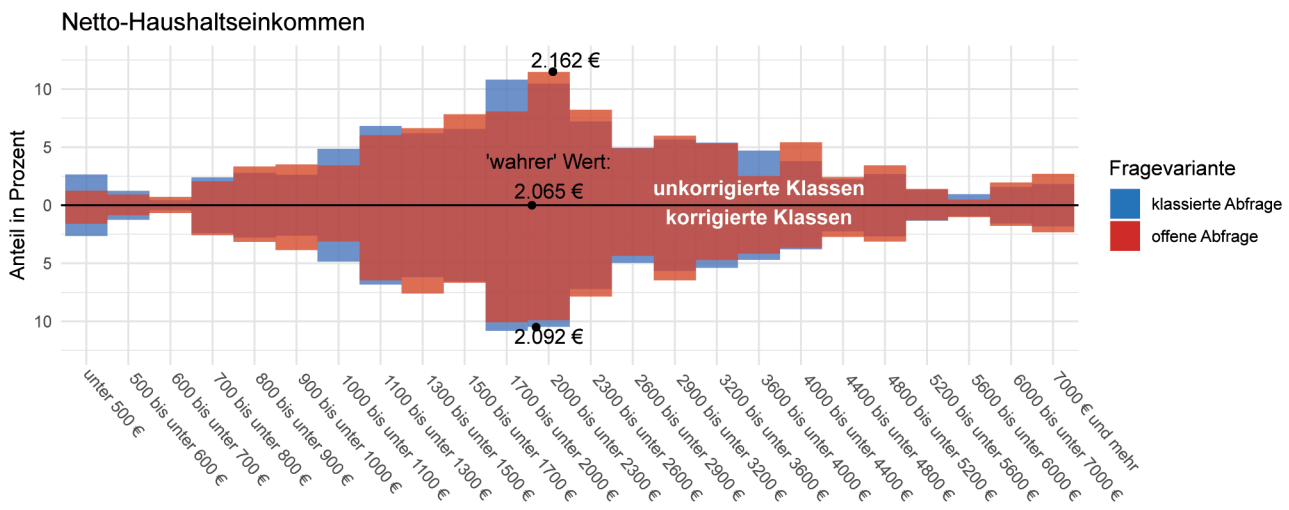
Nachdem nun die unterschiedlichen Ergebnisse, Herausforderungen und auch Implikationen aufgezeigt wurden, stellt sich die Frage, welches Instrument zur Erfassung des Einkommens geeigneter ist. Naturgemäß kann dies nur auf Basis der Forschungsfrage beantwortet werden. Stellt das Einkommen eine abhängige Variable dar, so bietet sich aufgrund der methodischen Vielfalt, welche im Zuge der Auswertung genutzt werden kann, eine offene Abfrage an. Auch wenn das Einkommen als ein zentraler Prädiktor zur Erklärung bestimmter Variablen fungieren soll, kann eine offene Abfrage sinnvoll sein, da der Zugewinn an Variabilität und die Erhöhung des Datenniveaus Vorteile bieten können, welche die höhere Antwortverweigerung ggf. rechtfertigen.

Stellt das Einkommen jedoch nur – und dies ist im Falle Kommunaler Bürgerumfragen häufig der Fall – eine mögliche Auswertungsperspektive dar und werden darüber hinaus häufig Cluster (arm vs. reich etc.) gebildet, so zeigt die aktuelle Analyse (Halbierung der Verweigerung bei klassierter Abfrage), dass eine klassierte Abfrage zur Senkung der Antwortverweigerungen zu empfehlen ist. Alternativ wäre auch die Erfassung der wirtschaftlichen Situation als Proxyvariable denkbar.

Um einen plastischeren Eindruck der Daten zu ermöglichen, stellt Abbildung 5 die Verteilungen getrennt nach Erhebungsmodus dar. Hierfür wurden die klassierten Daten in metrische Daten umgewandelt, sodass innerhalb einer Klasse jeder Wert nur einmal auftreten kann. Dies entspricht einer Adaption des Gleichverteilungsverfahrens zur Bestimmung des Medians beim Nettoäquivalenzeinkommen (Gerhardt et al. 2009: 7).

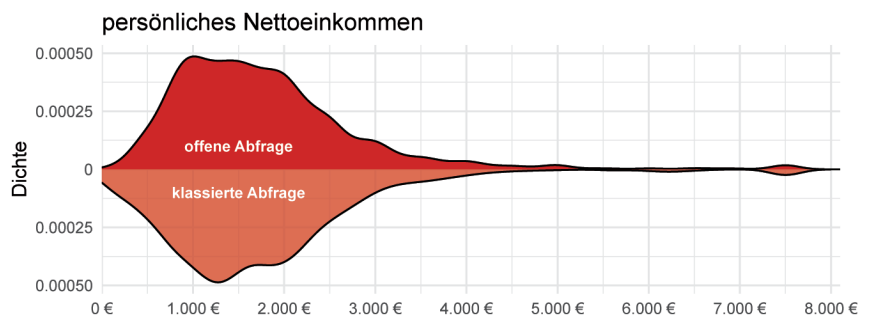
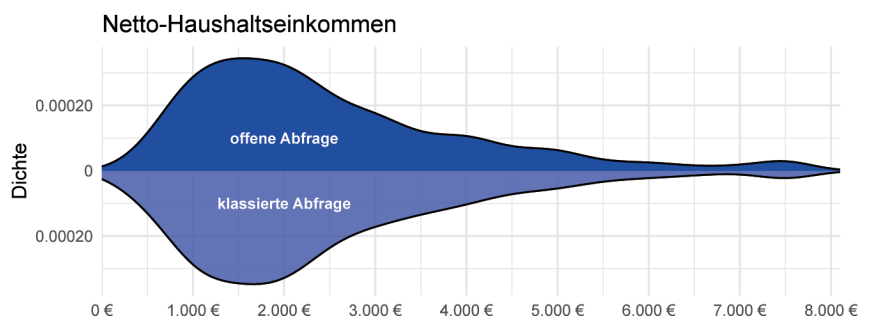
Dabei zeigt sich deutlich, dass die Verteilungen einen ähnlichen Verlauf haben, jedoch bestimmte Muster bzw. Feinheiten verloren gehen. Sofern die Klassen, insbesondere um den erwarteten Mittelwert und Median herum, möglichst fein gewählt werden, stellt dies jedoch eine ausreichende Datenbasis für eine verlässliche Schätzung des Medians dar.

Abb. 4: klassierte Einkommensverteilung nach Fragevariante



Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Abb. 5: Dichte der Einkommensverteilung differenziert nach Fragevariante



Lesehinweis: Die klassierten Daten wurden mittels Gleichverteilung in metrische Daten umgewandelt. Jeder Wert innerhalb einer Klasse kann dabei nur einmal besetzt werden. Werte größer 7.000 € wurden auf 7.500 € gesetzt.

Amt für Statistik und Wahlen Leipzig, Kommunale Bürgerumfrage 2021

Komplexere Analysen (wie z. B. Regressionsverfahren) auf Basis dieser umgewandelten Daten verbieten sich naturgemäß, es sei jedoch in diesem Beitrag zur Veranschaulichung gestattet:

Anhand der offen abgefragten Werte lassen sich mittels einfacher linearer Regression ca. 40 Prozent der Variabilität der Gesamtmiete anhand des Haushaltseinkommens erklären. Nimmt man die künstlich geschaffenen metrischen Werte, welche aus den klassierten erzeugt wurden, so sinkt die Varianzklärung auf 20 Prozent. Dieses Beispiel bestätigt zum einen die benannten Vorteile höherer Variabilität (der spitzen Werte), zum anderen das aufgeworfene Verbot einer solchen Analyse auf Basis künstlich geschaffener metrischer Daten.

Fazit

Zunächst bleibt festzuhalten, dass es sich bei der 2021 durchgeführten Kommunalen Bürgerumfrage um kein echtes experimentelles Design handelt: Hierfür wäre es z. B. erforderlich gewesen, dass in allen vier Fragebogenvarianten beide Erhebungsmodi jeweils zufällig den Befragten vorgelegt worden wären. Da die Kernaufgabe der Kommunalen Bürgerumfrage die Bereitstellung statistischer Daten zur Aufgabenerfüllung des Stadtrates und der Stadtverwaltung ist, stellt die in diesem Artikel vorgestellte Untersuchung nur ein Nebenprodukt dar.

Gleichwohl zeigen insbesondere die Regressionsmodelle (vgl. Tab. 2 und 3), dass kein reiner Fragebogeneffekt vorliegt, sondern die erhöhte Verweigerung in der Fragevariante begründet ist. Unabhängig davon gibt es weitere signifikante Prädiktoren für Item-Non-Response. Diese sind jedoch (teilweise) unabhängig von der Fragevariante.

Weiterhin sollte bei ähnlich angelegten Fragebögen – zumindest online – die Option für klassierte Abfragen gegeben werden, wenn die offene Abfrage verweigert wird (vgl. Schräpler 2004). Zusätzlich können sensibilisierende Begründungen bzw. ein nochmaliger Hinweis auf die Vertraulichkeit und Anonymität eingefügt werden.

Der vorliegende Beitrag bestätigt die bereits in der Vergangenheit in der Kommunalen Bürgerumfrage Leipzig beobachtete, aber nicht systematisch untersuchte Hypothese, nach der offene Einkommensabfragen zu einem geringeren

Rücklauf führen. Dass insgesamt die Antwortbereitschaft und damit die Rückläufe sinken, hat vielfältige Ursachen, deren systematische Analyse nicht Teil dieses Beitrags sind.

Als Fazit bleibt damit festzuhalten, dass die Vorteile der metrischen Erhebung die Nachteile der Verweigerungen aus Perspektive der Kommunalen Bürgerumfrage Leipzig nicht aufwiegen. In zukünftigen Erhebungen wird entsprechend zur Vermeidung des Item-Non-Response das Einkommen wieder klassiert erhoben.

- 1 Die Erfassung mittels offener Abfrage nimmt im Vergleich zur Erhebung mittels feiner Klassen deutlich weniger Platz im (Papier)Fragebogen ein. Da die Fragebogenlänge ein sehr wichtiges Kriterium darstellt, erscheint die Einsparung attraktiv.
- 2 Es handelt sich um kein echtes experimentelles Design. Hierfür wäre es z. B. erforderlich gewesen, dass in allen vier Fragebogenvarianten beide Erhebungsmodi jeweils zufällig den Befragten vorgelegt worden wären (vgl. Fazit).
- 3 Beispielsweise gab es Datensätze, bei denen eine allein lebende Person lediglich das persönliche, aber kein Haushaltseinkommen angab. Entsprechend wurde das Haushaltseinkommen als identisch angenommen. Jene Fehlwerte gehen nicht als Verweigerung in die Auswertung ein.
- 4 Die Fragebögen A–C sind hinsichtlich ihrer Stichprobenziehung repräsentativ für die Stadt Leipzig, zusätzlich wird eine Teilmenge, welche ebenfalls den Proportionen der Ortsteile in der Stadt Leipzig entspricht, dem Ortsteilfragebogen D entnommen. Die stadtweite Auswertung setzt sich somit aus den Fragebogenvarianten A–C sowie der aus D entnommenen Teilmenge zusammen.
- 5 Es wurden auch Modelle geprüft, bei denen für jede der Differenzierungsvariablen entsprechende Missings dummy-codiert einfließen. An den Befunden ändert dieses Vorgehen nichts, die Güte der Modelle war teilweise deutlich höher. Inhaltlich erscheint es jedoch fragwürdig Verweigerung eines Items durch Verweigerung eines anderen Items zu erklären.
- 6 Odds Ratios lassen sich wie folgt in Wahrscheinlichkeiten umrechnen: $OR / (1 + OR)$ siehe Grimes/Schulz (2008). Zu beachten ist dabei, dass sich jene Wahrscheinlichkeiten auf das entsprechende Stratum der Schätzung beziehen, in diesem Fall männlich, Hochschulabschluss, 18 bis 34 Jahre und erwerbstätig.
- 7 „Wahr“ bezieht sich dabei auf den rechnerisch richtigen Median. Der wahre Einkommenswert der Population bleibt wie in allen sozialwissenschaftlichen Erhebungen unbekannt. Die einfache Überführung der spitzen Werte in klassierte führt jedoch nachweislich zu einem falschen Ergebnis der Schätzung.

Literatur

Gerhardt, D. S. A., Habenicht, D. S. K., & Munz, D. S. D. E. (2009). Statistische Analysen und Studien, Band 58 Analysen zur Einkommensarmut mit Datenquellen der amtlichen Statistik. http://www.sozialberichte.nrw.de/sozialberichterstattung_nrw/grundlagen/statanalyse58.pdf

Grimes, D. A., & Schulz, K. F. (2008). Making sense of odds and odds ratios. *Obstetrics and gynecology*, 111(2 Pt 1), 423–426. <https://doi.org/10.1097/01.AOG.0000297304.32187.5d>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Schräpler, J.-P. (2004). Respondent Behavior in Panel Studies: A Case Study for Income Nonresponse by Means of the German Socio-Economic Panel (SOEP). *Sociological Methods & Research*, 33(1), 118–156. <https://doi.org/10.1177/0049124103262689>

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102(4), 989. <https://psycnet.apa.org/buy/2010-19349-001>

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York