AG Optimierung von Erhebungsdesigns bei Mixed Mode

Evaluierung und Dokumentation der Rekrutierungsexperimente bei der PUMA-Erhebung Q4/2016

Im Auftrag von der Plattform für Umfragen, Methoden und empirische Analysen (PUMA) und Statistik Austria

erstellt von
Dr. Alexander Seymer

Letzte Revision: 14. Juli 2017

Mein Dank gilt Statistik Austria und Matthias Till für die ausgezeichnete Zusammenarbeit. Die ausführlichen Diskussionen haben wesentlich zur Verbesserung beigetragen. Auch bei Martin Weichbold möchte ich mich für seinen Rat zu verschiedenen Detailfragen bedanken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Statistik Austria führt im Auftrag der Plattform für Umfragen, Methoden und empirische Analysen (PUMA) Rekrutierungsexperimente für zukünftige sozialwissenschaftliche Webbefragungen durch. Im Rahmen der Experimente soll eine möglichst optimale Form der Rekrutierung für zukünftige Befragungen identifiziert werden. Das Optimum ergibt sich in diesem Fall aus einer Abwägung von Qualität, Kosten und Umsetzbarkeit, wobei die Gewinnung qualitativ hochwertiger Daten als zentrales Kriterium hervorzuheben ist. Allerdings sei an dieser Stelle gleich angemerkt, dass mit höherem Aufwand in vielen Fällen die Daten zwar eine bessere Qualität aufweisen, dieser Zusammenhang allerdings einen asymptotischen Verlauf nimmt und deshalb bei der Betrachtung der Rekrutierungsexperimente das Qualitätskriterium allein nicht hinreichend ist.
Das in diesem Bericht vorgestellte Experiment ist das zweite von derzeit drei geplanten Experimenten. Das erste Experiment wurde im Frühjahr 2016 durchgeführt und fokussierte auf die Wirkung von Incentives, wozu vier unterschiedliche Incentives getestet wurden. Die Rekrutierung erfolgte aus der ausscheidenden Teilstichprobe des Mikrozensus heraus und damit standen Hintergrundinformationen zu jeder Person zur Verfügung, für welche eine detaillierte Analyse des Non-Response durchgeführt werden konnte. Die Stichprobenziehung des in diesem Bericht vorgestellten zweites Experiments erfolgte über Daten des Zentralen Melderegisters (ZMR). Neben der alternativen Stichprobenziehung wurden auch die Incentives entsprechend weiter ausdifferenziert, so dass im zweiten Experiment insgesamt sieben Incentivegruppen realisiert wurden.
Daraus leiten sich die zwei zentralen Teilbereiche für diesen Bericht ab. Zunächst muss das zweite Experiment unabhängig vom ersten auf die Qualität der Stichproben und die Wirkung der Incentives hin untersucht werden. Ähnlich wie im ersten Bericht werden detaillierte Betrachtungen des potentiellen Bias und der Kosten als zentrale Eckpfeiler der Evaluation herangezogen. Unabhängig vom Vergleich zum ersten Experiment werden zunächst Selektivität, Ausschöpfung und Kosten über die sieben unterschiedlichen Incentivegruppen analysiert. Im Anschluss werden die Ergebnisse zum ersten Bericht verglichen. Der Vergleich über die beiden Grunddimensionen Bias und Kosten sollte ein Vergleich über die unterschiedlichen Designs ermöglichen.
Da der erste Bericht (Seymer, 2017) bereits eine ausführliche Beschreibung der Literatur zu Incentives und Webbefragungen umfasst, wird in diesem Bericht die relevante Literatur nur als Ergänzung der Beschreibung des Experiments selbst angeführt. Nachdem das Design des Rekrutierungsexperimentes beschrieben wurde, folgt eine Herleitung relevanter Evaluationskriterien aus einer verkürzten theoretischen Darlegung zur Qualität von Umfragedaten, da auch diese bereits im ersten Bericht ausführlicher diskutiert wurde. Entlang der Kriterien erfolgt dann die Evaluation, bevor der Vergleich der beiden Experimente in die Schlussfolgerungen und Empfehlungen überführt wird.

Tabelle 1: Eigenschaften und geplante Bruttostichprobengröße der Incentivegruppen

Gruppe	Geldwert	Form	Zeitpunkt	Verpackung	N
A	0 EUR	kein Incentive			500
B	2 EUR	Sondermünze	unkonditional	Säckchen	450
C	2 EUR	Sondermünze	unkonditional	Folder	500
D	2 EUR + 5 EUR	Sondermünzen	unkonditional + konditional	Folder + Säckchen	500
E	5 EUR	Sondermünze	unkonditional	Säckchen	400
F	5 EUR	Sondermünze	konditional	Säckchen	400
G	10 EUR	Gutschein	konditional		1000

2 Experiment

Die Stichprobenziehung erfolgte einstufig aus dem Zentralen Melderegister (ZMR) zum Stichtag 1.10.2016 und umfasst die Bevölkerung Österreichs in der Altersgruppe zwischen 16 und 74 Jahren mit Hauptwohnsitz in privaten Haushalten (Till, 2017).[1] Dies entspricht einer Grundgesamtheit von 6.582.691 Menschen. Im ZMR stehen Informationen zu Bildung, Alter und Gemeinden zur Verfügung, welche für die Schichtung herangezogen wurden. Bildung wurde in die drei Gruppen differenziert: unbekannt oder maximal Pflichtschule (1), Lehre, berufsbildende Schule oder Matura (2) und Hochschule (3). Beim Alter wurden die 16- bis 30-Jährigen (1), die 31- bis 45-Jährigen (2), die 46- bis 60-Jährigen (3) und die 61- bis 74-Jährigen (4) in vier Gruppen unterschieden. Die Siedlungsdichte wurde über die Gemeinden mit einer Dreiteilung in niedrig (1), mittel (2) und hoch (3) realisiert. Daraus ergeben sich insgesamt 36 Schichten, aus denen jeweils der gleiche Auswahlsatz von 0,06 % ausgewählt wurde. Da für die Personen mit niedrigem Bildungsstand eine geringe Ausschöpfung erwartet wurde, ist der Auswahlsatz für diese Schichten um 50 % auf 0,09 % angehoben worden (vgl. Tabelle 2).
Daraus resultierte eine Bruttostichprobe von 4.356 Adressen, welche randomisiert auf sechs Gruppen (A-F) mit 500 Adressen und eine Gruppe (G) mit 1.356 aufgeteilt wurden, um die sieben Incentivegruppen abzubilden. Jede der sieben Stichproben wurde in zwei Versandtranchen aufgeteilt. Für die zweite Versandtranche wurden basierend auf den Rücklaufzahlen zur Kostenreduktion und in Absprache mit PUMA die ursprüngliche Stichprobengrößen von vier der sieben Stichproben angepasst. Dementsprechend wurden nur 3.750 Adressen für die Versendung genutzt. Im Detail erfolgte die Verkleinerung der zweiten Tranche für die Gruppe B von 250 auf 200 Adressen, für Gruppe E und F jeweils von 250 auf 150 Adressen und für die Gruppe G von 1.356 auf 1.000 Adressen.
Meist werden Incentives nach Zeitpunkt und Form des Incentives differenziert (vgl. Ernst Stähli & Joye, 2016). In den sieben Incentivegruppen dieses Experiments werden sowohl der Zeitpunkt als auch die Form des Incentives getestet (vgl. Tabelle 1). Beim Zeitpunkt wurden die klassischen Designs von unkonditionalem oder bedingungslosem Incentive direkt mit dem Anschreiben (Gruppe B, C, E) und das konditionale Incentive nach der Teilnahme (Gruppe F, G) realisiert. Bei der Form des Incenti-

[1] Eine detailliertere Darstellung der Feldphase, als hier mit Blick auf die Ausschöpfung nötig ist, erfolgt im Feldbericht der Statistik Austria (Till, 2017)

Tabelle 2: Verteilung der Stichprobe (inkl. „Reserveadressen“) auf Schichten (vgl. Tabelle 1 Till, 2017)

	Bildung	Alter	Siedlungsdichte	Grundgesamt	Bruttostichprobe	Auswahlsatz
1	1	1	1	235.390	207	0,09
2	1	1	2	203.193	179	0,09
3	1	1	3	230.322	203	0,09
4	1	2	1	138.832	122	0,09
5	1	2	2	93.123	82	0,09
6	1	2	3	76.178	67	0,09
7	1	3	1	124.979	110	0,09
8	1	3	2	109.974	97	0,09
9	1	3	3	145.191	128	0,09
10	1	4	1	84.456	74	0,09
11	1	4	2	95.433	84	0,09
12	1	4	3	152.410	134	0,09
13	2	1	1	287.204	168	0,06
14	2	1	2	228.608	134	0,06
15	2	1	3	325.677	191	0,06
16	2	2	1	272.370	160	0,06
17	2	2	2	320.010	188	0,06
18	2	2	3	496.717	291	0,06
19	2	3	1	312.452	183	0,06
20	2	3	2	408.742	240	0,06
21	2	3	3	608.526	357	0,06
22	2	4	1	213.070	125	0,06
23	2	4	2	235.803	138	0,06
24	2	4	3	303.556	178	0,06
25	3	1	1	59.216	35	0,06
26	3	1	2	24.533	14	0,06
27	3	1	3	26.623	16	0,06
28	3	2	1	177.486	104	0,06
29	3	2	2	95.203	56	0,06
30	3	2	3	90.428	53	0,06
31	3	3	1	125.045	73	0,06
32	3	3	2	84.847	50	0,06
33	3	3	3	74.488	44	0,06
34	3	4	1	55.350	32	0,06
35	3	4	2	35.875	21	0,06
36	3	4	3	31.381	18	0,06
Gesamt				6.582.691	4.356	0,07

ves wurde eine zusätzliche Unterscheidung vorgenommen. Neben dem Geldwert des Incentives variierte auch die Form, indem entweder eine Sondermünze oder ein Gutschein eingesetzt wurde, und auch die Art der Verpackung. Als Geldwerte wurden 2 EUR in Form einer Sondermünze zum 200-jährigen Bestehen der Österreichischen Nationalbank, eine 5 EUR Sondermünze aus Kupfer mit Albrecht Dürer's Feldhase und ein universell einsetzbarer 10 EUR Einkaufsgutschein der Firma Edenred gewählt. Als Verpackung für die Münzen wurde entweder ein rotes Säckchen, welches händisch an die Schreiben angeheftet werden musste, oder ein speziell für die Statistik Austria produzierter Informationsfolder mit der eingeschweißten Münze verwendet. Der Informationsfolder kam nur in Gruppe C und D zur Anwendung.
Im ersten Rekrutierungsexperiment wurden vier Incentivegruppen unterschieden. Dabei handelte es sich generell um unkonditionale Incentives, welche bei der telefonischen Rekrutierung als bedingungslose Incentives angekündigt und mit dem Aviso-Brief den Befragten ausgehändigt wurde. Eine Gruppe erhielt als Sachincentive die Broschüre Zahlen-Daten-Fakten, die anderen drei Gruppen waren in Geldwert und Form identisch zu den Gruppen B, E und G des zweiten Experiments, wobei der Gutschein ebenfalls unkonditional vergeben wurde. Dementsprechend ist streng genommen nur ein Vergleich von Gruppe B und E mit den entsprechenden Gruppen des ersten Experiments möglich, wobei im ersten Experiment die 2 EUR Gruppe die besten Ergebnisse lieferte.
Die Versendung der ersten Tranche von 2.176 Anschreiben erfolgte am 10.11.2016, wobei durch eine technische Panne die Fragebögen erst am 11.11.2016 gegen 12 Uhr freigeschaltet werden konnten. Außerdem wurde versehentlich der Fragebogen der ersten Piloterhebung freigeschaltet. Dieser Fehler wurde am 22.11.2016 behoben. Bis dahin hatten 206 Personen mit der Beantwortung des falschen Fragebogens begonnen, wobei 80 Personen davon auch den richtigen Fragebogen ausgefüllt hatten. Die restlichen 126 Fragebögen wurden als fehlgeleitet eingestuft und über ein gezieltes Erinnerungsschreiben zum Ausfüllen des korrekten Fragebogens konnte die Zahl der fehlgeleiten Fragebögen auf 69 reduziert werden. Zusätzlich waren 26 Anschreiben der ersten Tranche nicht zustellbar, und deshalb reduziert sich die berücksichtigte Bruttostichprobe auf 2.081 (vgl. Till, 2017).
Trotz der technischen Problem zum Beginn des Experiments übertrafen die Rücklaufquoten deutlich die Erwartungen. Aufgrund der Incentivekosten war deshalb auch mit einer Kostenüberschreitung zu rechnen, wenn das Experiment wie geplant mit der vollen Stichprobe fortgesetzt werden würde. Deshalb wurde durch Statistik Austria in Abstimmung mit PUMA die zweite Versandtranche auf 1.574 Stichprobenpersonen reduziert. Aus der zweiten Versandtranche wurden 23 als nicht zustellbar retourniert. Die zweite Tranche wurde am 6.12.2016 versandt und zur Steigerung der Ausschöpfung wurden 12 Tage nach der Aussendung der ersten Versandtranche und 8 Tage nach Ausgang der zweiten Versandtranche Erinnerungspostkarten ausgesendet. Die letzte registrierte Bearbeitung eines Fragebogens erfolgte am 26.1.2017 und mit dem 31.1.2017 wurde die Feldphase offiziell beendet und der Fragebogen deaktiviert.
Die unzustellbaren Anschreiben und die falsch ausgefüllten Fragebogen verringerten die Bruttostichprobe um 118 Fälle, so dass im Zeitraum des Experiments von einer bereinigten Bruttostichprobe von 3.632 Personen ausgegangen werden kann (vgl. Tabelle 3). Insgesamt konnten 721 gültige Fragebögen realisiert werden, wobei alle

Tabelle 3: (bereinigte) Bruttostichprobe nach Versandtranche

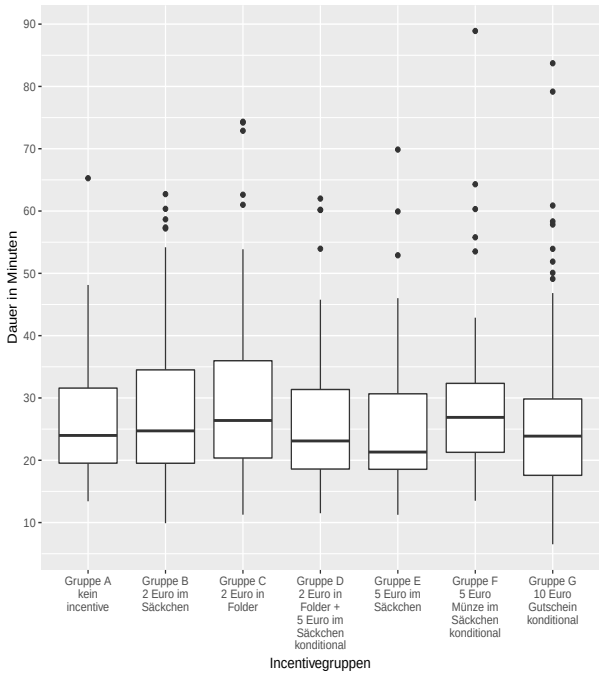
Incentivegruppe	Bruttostichprobe	Unzustellbar	Falsch	Insgesamt
Gruppe A	235	5	10	250
Gruppe B	238	3	9	250
Gruppe C	241	2	7	250
Gruppe D	233	3	14	250
Gruppe E	237	5	8	250
Gruppe F	243	2	5	250
Gruppe G	654	6	16	676
1. Tranche	2081	26	69	2176
Gruppe A	244	6		250
Gruppe B	198	2		200
Gruppe C	247	3		250
Gruppe D	248	2		250
Gruppe E	148	2		150
Gruppe F	147	3		150
Gruppe G	319	5		324
2. Tranche	1551	23		1574

Fragebögen mit 13 und mehr fehlenden Antworten als abgebrochene oder ungültige Fragebögen eingestuft wurden (vgl. Tabelle 8, S. 18)

2.1 Web-Befragung

Die Web-Befragung selbst bestand genauso wie die erste Piloterhebung aus vier Modulen. Das Modul zu den soziodemographischen Variablen orientierte sich hierbei an der ersten Piloterhebung und Variablen, welche für die erste Piloterhebung über den Mikrozensus zur Verfügung standen. Die anderen drei Module behandeln das subjektive Wohlbefinden, die Persönlichkeit und politische Teilhabe sowie das Verdienen von wohlfahrtstaatlichen Leistungen und Immigration. Die Vorgabe für alle Module war eine Bearbeitungszeit von 5 Minuten. Trotzdem variierte die Anzahl der Frageitems deutlich. Das Modul zum subjektiven Wohlbefinden umfasste 15 Fragen die auf einer identischen 11er Skala zu beantworten waren. Das Modul zur Persönlichkeit und politischen Teilhabe umfasst 13 Teilkomponenten die sich in 65 Items untergliedern und unterschiedlichen Skalen besitzen. Das Modul zum Verdienst von wohlfahrtstaatlichen Leistungen und Immigration war als Vignettendesign angelegt und es wurden den befragten Personen fünf Vignettenpaare vorgelegt, wobei jeweils drei Fragen beantwortet werden sollten.
Die Abbildung 1 verdeutlicht, dass die Dauer der Web-Befragung für eine große Zahl der TeilnehmerInnen mehr Zeit erforderte als geplant. Dies ist im Hinblick auf allfällige Wiederholungsbefragungen kritisch zu hinterfragen. Ein statistisch signifikanter

Abbildung 1: Dauer der Web-Befragung nach Incentivegruppen



Zusammenhang zwischen Dauer der Web-Befragung und der Bereitschaft zur wiederholten Teilnahme, welche über die Weitergabe der E-Mailadresse erfasst wurde, besteht jedoch nicht.

2.2 Fragebogenabbruch

Als Befragungsabbruch wird verstanden, wenn die adressierte Person, den Fragebogen öffnet, aber nicht bis zum Ende beantwortet. Hierbei können mehrere Formen von Abbrüchen unterschieden werden (vgl. Callegaro, Manfreda & Vehovar, 2015, S. 138f). Einerseits kann die Teilnahme gleich zu Beginn oder im Rahmen der Einleitung oder andererseits im späteren Verlauf abgebrochen werden. Der Abbruch gleich zu Beginn lässt sich auf andere Ursachen zurückführen als der Abbruch im späteren Verlauf. Ein Abbruch zu Beginn des Surveys kann wesentlich durch die Eigenschaften des Fragebogens (bspw. die Darstellungsform) oder situative Rahmenbedingungen (bspw. Zeitdruck) hervorgerufen werden. Der Abbruch während der Befragung ist unmittelbarer auf die spezifischen Eigenschaften des Fragebogens (bspw. Länge, Inhalt, Art der Fragestellung) oder Umstände der befragten Person (bspw. Ermüdung oder Ablenkungen) zurückzuführen.
Im Kontext von Webbefragungen ist das Gerät, auf dem die Beantwortung erfolgt, ein wesentliches Kriterium für einen möglichen Abbruch. Neben dem PC ist auch die Teilnahme per Smartphone oder Tablet möglich. Drewes (2014) zeigen in einer Untersuchung zum Einfluss von Smartphones auf Panel Onlinesurveys, dass der Zeitraum zwischen Einladung und erstem Zugriff auf den Fragebogen bei Smartphones geringer ist als bei PCs. Gleichzeitig benötigt die Bearbeitung des Fragebogens am Smartphone mehr Zeit. In der Studie von Drewes (2014) nutzen aber nur 4 % aller

erfolgreich Befragten das Mobiltelefon.
In der zweiten Piloterhebung wurden 771 Zugriffe auf den Fragebogen registriert und 721 vollständige Fragebogen erfasst. Damit brachen 6,5 % aller Personen, die den Fragebogen begonnen hatten, die Befragung ab. Von den 771 Zugriffen erfolgten 46 (6,0 %) per Mobiltelefon, 45 (5,8 %) per Tablet, 498 (64,6 %) per PC und 182 (23,6 %) konnten keinem Gerät zugeordnet werden. Auffallend ist die hohe Abbruchquote von 28,3 % unter den Mobiltelefonnutzern und 13,3 % unter den Tablet-Nutzern. Bei den Nutzern von PCs fällt die Abbruchquote mit 4,8 % deutlich geringer aus und aufgrund der geringen Abbruchquote von 3,8 % unter den unbekannten Geräten, wurden für diese Fragebögen vermutlich mehrheitlich PCs verwendet. Die Anteilswerte stimmen durchaus mit der Literatur überein und sind bei den Mobiltelefonen auch auf eine längere Antwortzeit zurückzuführen. Ein Mittelwertvergleich für alle vollständigen Fragebögen zeigt einen signifikanten (F=11,596; df=3; p=0,000) Unterschied. Mobiltelefone sind mit einer durchschnittlichen Bearbeitungszeit von 35,4 min. deutlich länger mit dem Beantworten beschäftigt als Tablet-Nutzer (29,5 min.), PC-Nutzer (27,3 min.) oder die Nutzer, denen kein bestimmtes Gerät zugewiesen werden kann (23,9 min.).
Die Schlussfolgerung von Drewes (2014), dass nicht die technische Barriere der schlechteren Darstellbarkeit das größere Problem für Befragung über Mobiltelefone darstellt, sondern die erhöhte Beantwortungsdauer, erscheint im Zusammenhang dieser Ergebnisse durchaus plausibel. Gerade Mobiltelefone werden auch in Situationen verwendet, wo die notwendige Ruhe und Zeit zur Beantwortung eines Fragebogens fehlt und damit die Chance für Ablenkungen und Unterbrechungen steigt.
Eine Überprüfung auf einen Zusammenhang zwischen Abbrechen und Bildung wies signifikante Ergebnisse aus (χ^2=7,764; df=2; p=0,021). Unter den Abbrechern ist der Anteil an Personen mit niedriger Bildung deutlich erhöht und der mit hohem Bildungsniveau deutlich geringer. Interessanterweise ist für die Personen mit mittlerem Bildungsniveau, welche mit 438 Personen die größte Gruppe stellen, kein Unterschied zwischen Abbrechen und Teilnahme zu erkennen (vgl. Tabelle 24, S. 36).
Wenn man die Abbrüche über Bildung und genutztem Endgerät vergleicht (vgl. Tabelle 25 und Tabelle 26, S. 37), dann stellt sich heraus, dass Befragte mit niedrigem Bildungsniveau deutlich häufiger das Mobiltelefon zur Beantwortung nutzen. Scheinbar hat auch das Nutzungsverhalten der Geräte einen Einfluss auf die Ausschöpfung und könnte Teile der geringeren Ausschöpfung erklären.[2]

2.3 Wiederholungsbefragung

Die Rekrutierungsexperimente dienen auch der Identifikation der optimalen Strategie zur Gewinnung von TeilnehmerInnen an einem Onlinepanel. Deshalb wurde in der zweiten Piloterhebung um die Bereitstellung einer E-Mailadresse gebeten, falls die befragte Person an der Teilnahme einer weiteren Befragung interessiert ist. Damit stellt sich auch die Frage, ob Incentives Einfluss auf die Teilnahme an weiteren Befragungen haben. Die Forschung zum Einsatz von Incentives konzentriert sich mehrheitlich

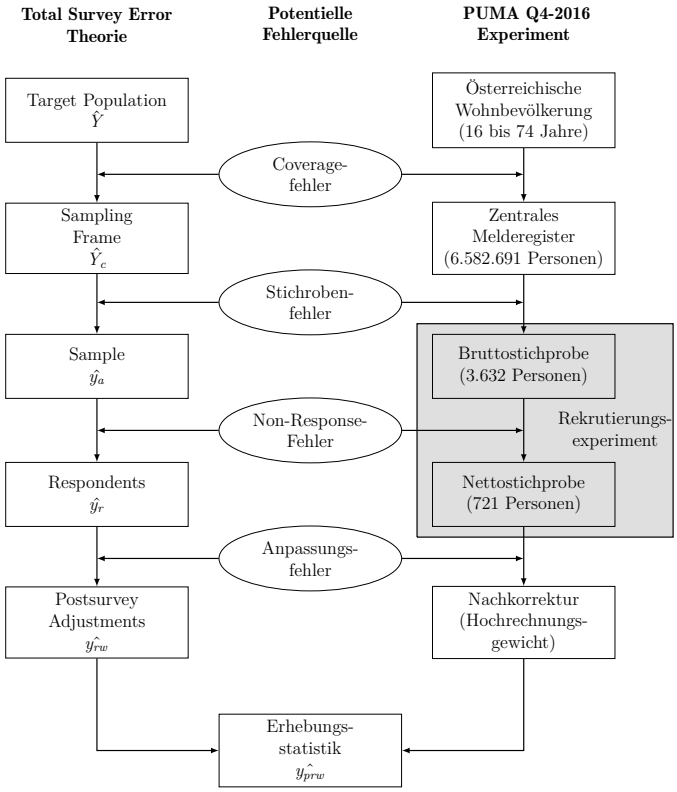
[2]Ein Test auf Unabhängigkeit in einer dreidimensionalen Kreuztabelle mit Bildung, Art des Gerätes und Abbruch weist ein signifikantes Ergebnis aus (LR-χ^2=12,999; df=6; p=0,043). Dieser Wert ist aber aufgrund der geringen Fallzahlen der Abbrecher nur als Trend interpretierbar.

auf die Gewinnung neuer TeilnehmerInnen und diesbezüglich unterstützen die Ergebnisse über die letzten Jahrzehnte den Einsatz von unkonditionalen Incentives als die Strategie mit den besten Ergebnissen in der Erhöhung der Ausschöpfung und der Datenqualität (vgl. Callegaro et al., 2014). Ein negativer Effekt von Incentives könnte die sinkende Teilnahmebereitschaft sein, da die Befragten sich an Incentives gewöhnen könnten. Die Befunde für solche Effekte sind uneindeutig und bisher umstritten (Singer & Ye, 2013, S. 133f). Inwieweit diese Ergebnisse zu den Incentiveeffekten bei der Rekrutierung für Querschnittsstudien auch für Panelstudien gelten, ist weniger intensiv erforscht (siehe bspw. Blom et al., 2016).
Laurie und Lynn (2009) untersuchen den Einsatz in unterschiedlichen englisch- und deutschsprachigen Panelstudien und kommen zu dem Ergebnis, dass für die Erstrekrutierung die Unterschiede in den Effekten der Incentives zwischen Panelstudien und Querschnittsstudien eher gering sind. Ihre Analysen deuten darauf hin, dass Incentives besonders den Attrition Bias positiv kompensieren könnten. Allerdings konnten die Autoren keine eindeutigen Belege vorweisen. Grundsätzlich dürfen Incentives auch keinesfalls als einzige Strategie zur Erhöhung der Ausschöpfung oder Datenqualität betrachtet werden. Schoeni, Stafford, Mcgonagle und Andreski (2013) diskutieren die Strategien von unterschiedlichen Panelstudien in den USA und zeigen, dass die Höhe des Incentives und die Länge des Fragebogens die sehr hohen Ausschöpfungen von jenseits der 90 % nur geringfügig beeinflussen (vgl. Schoeni et al., 2013, Figure 3, S. 75). Vielmehr erreichen die Surveys die Befragten über die Anzahl der Kontaktversuche, Nachverfolgungsstrategien, die Anwendung von unterschiedlichen Befragungsmodi und die Steigerung der wahrgenommenen Qualität des Surveys über Meinungsführer. Diese Ergebnisse stimmen mit den Schlussfolgerungen von Blom et al. (2016) überein und besonders die wahrgenommene Qualität des Surveys ist im Kontext der Pilot II Erhebung interessant für die Beurteilung der Gruppe C, wo der Folder explizit mit dieser Zielstellung getestet wurde.

3 Evaluationskriterien

Dem Ansatz der Evaluation der ersten Piloterhebung (Seymer, 2017) folgend soll sich die Evaluation grob an der Total Survey Error Theory (Groves et al., 2009; Lyberg & Weisberg, 2016) orientieren. Die Abbildung 2 verdeutlicht die Parallelen zwischen Theorie und Rekrutierungsexperiment. Die Unterscheidung in die potentiellen Fehlerquellen dient an dieser Stelle einer ganzheitlichen Betrachtung aller theoretisch möglichen Fehlerquellen. Es wird im Verlauf der Diskussion schnell deutlich, dass im vorliegenden Fall einige Fehlerquellen nicht plausibel sind oder über das Design ausgeschlossen werden konnten.
Der Coveragefehler erfasst die Diskrepanz zwischen Grundgesamtheit und Auswahlgesamtheit, wobei in diesem Kontext zwischen Undercoverage und Overcoverage unterschieden werden kann. Als Undercoverage werden die Personen gezählt, welche zwar in der Grundgesamtheit sind, aber nicht in der Auswahlgesamtheit. Beim Overcoverage sind Personen in die Auswahlgesamtheit aufgenommen worden, obwohl sie nicht in der Grundgesamtheit vorkommen.
Über den Stichprobenziehungsfehler werden Probleme im Prozess der Stichproben-

Abbildung 2: Total Survey Error Theorie im Abgleich zur zweiten Piloterhebung



ziehung selbst erfasst. Da der Auswahlrahmen normalerweise nicht komplett erhoben wird, sondern eine kleinere Stichprobe aus allen erreichbaren Personen ausgewählt wird, kann in diesem Schritt eine Verzerrung auftreten. Hier werden im Wesentlichen die Fragen der Erreichbarkeit diskutiert.
Der Non-Response-Fehler ist für die Zielstellung der Rekrutierungsexperimente der zentrale Punkt. Im Rahmen des Non-Response-Fehlers wird die Frage erörtert, ob die Personen, welche die Teilnahme verweigern, die Stichprobe systematisch verzerren. Die Rekrutierungsexperimente versuchen hier eine optimale Strategie zu identifizieren, wie Verzerrungen durch Non-Response möglichst geringgehalten werden können. Auch die Frage der Kosten ist hier relevant, denn jede Person, welche die Teilnahme verweigert, verursacht bereits Kosten ohne, dass Informationen gewonnen werden konnten.
Der Vergleich der Incentivegruppen wird wesentlich im Rahmen des Non-Response-Fehlers und dessen potentieller Veränderung diskutiert. Unabhängig davon werden die Incentivegruppen auch an anderer Stelle immer wieder herangezogen, um etwaige Zusammenhänge zu beleuchten.
Trotz großer Bemühungen können nicht alle potentiellen Verzerrungen vermieden werden. Im Mikrozensus war die Teilnahme obligatorisch, so dass Verzerrungen im Vergleich zu freiwilliger Befragungsteilnahme verringert werden können (vgl. Gumprecht & Oismüller, 2013). Bei der Befragung in diesem Experiment erfolgte die Rekrutierung ausschließlich über Brief und damit besteht relativ wenig Kontrolle über finale Selektion der befragten Person. Um die gewonnen Ergebnisse trotzdem möglichst gut

Tabelle 4: Überblick Evaluationskriterien

	Coveragefehler	Stichprobenfehler	Non-Response-Fehler	Anpassungsfehler	Kosten
Kennzahlen	Theoretische Diskussion	Kreuztabelle	Ausschöpfung, Beschreibung Bias, R-Indikator, Log. Regression	Designeffekt, Variationskoeffizient, Quartilsdispersionskoeffizient	Ausschöpfung
ZMR-Variablen	Alter, Geschlecht, Siedlungsdichte, Bildung				
PUMA-Variablen				Zufriedenheit mit dem Leben, Teilnahme Nationalratswahlen, Vignette: finanzielle Hilfe vom Staat	

nutzen zu können, werden Anpassungsgewichte, welche die Eigenschaften der Erhebung berücksichtigen und Hochrechnungsgewichte berechnet, um die Daten perfekt entlang unterschiedlicher Variablen auf die Grundgesamtheit umzulegen.
Nachdem die theoretisch möglichen Fehlerquellen diskutiert worden sind, wird noch über die Kosten reflektiert. Hier wird ähnlich wie im ersten Bericht eine Schätzung standardisiert auf eine fixe Nettostichprobe von n=1000 erstellt, die auf der incentivespezifischen Ausschöpfung und den Incentivekosten beruht. Damit sollte vor allem ein grober Vergleich zu den Kosten der unterschiedlichen Incentivegruppen aus dem ersten Bericht möglich sein.

3.1 Kennzahlen und Variablen

Im Bericht zur ersten Piloterhebung wurden sowohl die Kennzahlen als auch die Variablen explizit diskutiert und in der Auswahl begründet. Tabelle 4 stellt die Verwendung für diesen Bericht dar. In der Aufzählung zeigt sich der Fokus auf die Betrachtung des Non-Response-Fehlers. Die Ausschöpfung wird über das Verhältnis von Netto- zu Bruttostichprobe bestimmt, wobei falsch ausgefüllte Fragebögen (69), Retoursendungen (49), Abbrüche und ungültig Fragebögen (50) alle als Non-Response behandelt werden. In den AAPOR Empfehlungen (AAPOR, 2016) entspricht dies Response Rate 1, der sogenannten minimalen Responserate.
Die Ausschöpfung selbst ist kein hinreichendes Kriterium für die Qualität eines Surveys, denn die Verzerrung variiert über jede Variable unterschiedlich (vgl. Groves et al., 2009). Theoretisch kann auch eine sehr geringe Ausschöpfung zu einer sehr repräsentativen Stichprobe führen. Um den Einfluss unterschiedlicher Variablen auf die Ausschöpfung zu untersuchen, wird eine gewichtete logistische Regression auf die Teilnahme an der Befragung berechnet. Als Gewichtungsfaktor wird das Designgewicht herangezogen, welches das Oversampling der Personen mit niedriger Bildung korrigiert. Der R-Indikator ist als standardisiertes Maß der Repräsentativität konzipiert (de Heij, Schouten & Shlomo, 2010), welches auf einem gewichteten logistischen Regressionsmodell beruht und für den Vergleich über unterschiedliche Populationen

besser geeignet ist als etwa das Pseudo-R^2 nach Nagelkerke. Dazu werden neben dem R-Indikator selbst auch ergänzend der maximalen Bias und der maximale Kontrast als Maßzahlen der Verzerrungen hinzugezogen (vgl. de Heij et al., 2010, S. 17ff). Der Designeffekt nach Lumley (2016) ermöglicht den Vergleich der Stichprobenziehungseffekte komplexer Designs gegen eine einfache Zufallsauswahl. Bei der Berechnung des Designeffekts wird die Anpassungsgewichtung, welche als Korrektur für die Stichprobenziehungseffekte eingeführt wird, genutzt. Deshalb kann der Designeffekt als Kenngröße auch als Maßzahl für den Anpassungsfehler herangezogen werden.

4 Evaluation

Zunächst wird das Rekrutierungsexperiment für sich betrachtet und die möglichen Vergleiche innerhalb der zweiten Piloterhebung vorgenommen. Dazu zählen neben der Betrachtung der unterschiedlichen Fehlerquellen über die Erhebung die Unterscheidung der Incentivegruppen. In der ersten Piloterhebung sind vier sehr unterschiedliche Incentives erprobt worden. Mit der zusätzlichen Ausdifferenzierung in sieben Incentivegruppen können auch Teilaspekte der Incentives verglichen werden (vgl. Tabelle 1, S. 5). Insgesamt können Geldwert, Form, Zeitpunkt und Verpackung unterschieden werden. Die Geldwerte und Form konnten schon in der ersten Pilotstudie verglichen werden, allerdings waren in der Erhebung alle Incentives bedingungslos und es fehlte eine Gruppe ohne Incentive. In der zweiten Welle wurden nun auch konditionale Incentives eingesetzt, so dass auch ein Vergleich zu den bedingungslosen Incentives möglich ist. Außerdem wurde auch die Verpackung der Incentives variiert. In dieser zweiten Erhebung kam ein speziell entworfener Folder als alternative Verpackung hinzu. Die Diskussion dieser vier Unterschiede in den Incentivegruppen wird parallel zur Betrachtung der potentiellen Fehlerquellen und primär im Rahmen der Diskussion zum Non-Response-Fehler vollzogen.

4.1 Coverage-Fehler

Als Grundgesamtheit wurde die in Privathaushalten lebende Wohnbevölkerung Österreichs im Alter von 16 bis 74 Jahren definiert. Mit dem ZMR als Auswahlgrundlage ist eigentlich kein Overcoverage möglich, denn jede verfügbare Adresse ist per Definition Teil der Grundgesamtheit. Ein verwandtes Problem zu Overcoverage ist die Situation, dass nicht die adressierte Person an der Befragung teilnimmt, sondern eine zufällig anwesende Person, welche nicht aus der Grundgesamtheit stammt. Alternativ kann Overcoverage bei fehlenden Ummeldungen auftreten. Dieses Problem ist durchaus nicht zu unterschätzen, so hat der Zensus 2011 in Deutschland eine Diskrepanz von 1,5 Millionen Einwohnern zwischen den Fortschreibungen auf der Basis der letzten Zensus Daten und den durch den Zensus 2011 erhobenen Daten nachgewiesen (vgl. Ämter des Bundes und der Länder, 2011, S. 11).
Undercoverage dürfte im Kontext der PUMA Erhebung über ZMR Stichproben eine relativ geringe Rolle spielen, denn dabei kann es sich nur um Phänomene wie verzögerte Wohnsitzanmeldungen, „illegale“ Einwanderung oder Leben am Zweitwohnsitz handeln. Aufgrund der engen Verzahnung des Hauptwohnsitzes mit dem täglichen

Tabelle 5: Unabhängigkeitstests für Stichprobenziehungsfehler durch Bereinigung der Bruttostichprobe (118 vs. 3632)

Variable	Verfahren	Statistik	df	p.Wert
Altersgruppen laut Bericht	Pearson's Chi-squared test	3,784	5	0,581
Geschlecht	Pearson's Chi-squared test with Yates' continuity correction	0,009	1	0,923
Alter ungruppiert	Welch Two Sample t-test	-0,311	124	0,756
Bildung laut Schichtung	Pearson's Chi-squared test	0,817	2	0,665
Altersgruppen laut Schichtung	Pearson's Chi-squared test	1,435	3	0,697
Siedlungsdichte laut Schichtung	Pearson's Chi-squared test	11,854	2	0,003

Leben (Bankkonto, Arbeitgeber etc.), ist es aber unplausibel, dass Undercoverage von großer Bedeutung ist und außerdem wird das Problem über die Definition der Grundgesamtheit schon ausgeschlossen. Man könnte aber argumentieren, dass alle in Österreich lebenden Personen berücksichtigt werden sollten, dann würden Personengruppen wie Obdachlose ebenfalls erfasst werden müssen. In diesem Fall wäre ein Undercoveragebias vorhanden.
Diese theoretische Diskussion sollte nur kurz verdeutlichen, dass Coverage kein wesentliches Problem für eine ZMR-Stichprobe darstellt, wenn die von PUMA definierte Grundgesamtheit zugrunde gelegt wird. Bei einer anders definierten Grundgesamtheit sind auch für ZMR-Stichproben Coverageprobleme möglich.

4.2 Stichprobenziehungsfehler

Auch der Stichprobenziehungsfehler ist für dieses Experiment nur von nachgeordneter Relevanz, denn die Auswahl aus dem ZMR erfolgte randomisiert. Auch die zweite Auswahl aufgrund der nachträglichen Reduktion der zweiten Tranche erfolgte randomisiert, so dass die 3.750 Adressen, welche letztlich verwendet wurden als Zufallsauswahl betrachtet werden können. Von diesen 3.750 Adressen werden aber nur 3.632 als bereinigte Bruttostichprobe weiter berücksichtigt und die 118 fehlenden Adressen sollen hier als Stichprobenziehungsfehler diskutiert werden. Nach APPOR Empfehlungen sind Rücksendungen auch als Non-Response interpretierbar, allerdings soll an dieser Stelle die technische Panne bezüglich des falschen Fragebogens und die Rücksendungen vom Non-Response getrennt werden, denn diese beiden Personengruppen hatten keine Möglichkeit an der Befragung teilzunehmen. Und die Tatsache, dass eine Teilnahme nicht möglich war, soll als Stichprobenziehungsfehler interpretiert werden und die Berücksichtigung einer bereinigten Bruttostichprobe für die Non-Response-Analyse ermöglichen.
Eine zusätzliche Fehlerquelle, die allerdings nicht kontrollierbar ist, stellt die unbekannte Anzahl an unretournierten Anschreiben dar, welche den Adressaten nicht erreicht haben. Zum Umfang dieses Fehlers gibt es keine gesicherte Informationsgrundlage, deshalb kann dies nicht in die Berechnung der Ausschöpfung einfließen. Grundsätzlich würde dies letztlich zu einer Erhöhung der Ausschöpfung im Verhältnis zur bereinigten Bruttostichprobe führen. Das Gesamtdesign der Erhebung würde sich

Tabelle 6: Gewichtete Mittelwerte, Standardfehler und Designeffekte nach Lumley einzelner Variablen aus der PUMA Befragung

Variable	Mittelwert	Standardfehler	Designeffekt
Zufriedenheit mit dem Leben (0 = unzufrieden, 10 = zufrieden)	7,45	0,08	1,19
Teilnahme Nationalratswahlen (0 = sicher nicht, 10 = sicher ja)	8,85	0,10	1,16
Vignette: finanzielle Hilfe vom Staat (1 = keinesfalls, 7=jedenfalls)	5,91	0,05	1,14

dadurch aber nicht verbessern, denn der Fehler müsste sowohl in den Kosten als auch bei zukünftigen Erhebungen mitberücksichtigt werden.
In Tabelle 5 sind die Ergebnisse von Unterschiedstests zwischen den 118 Personen, die nicht teilnehmen konnten, und den 3.632 Personen aus der bereinigten Bruttostichprobe dargestellt. Da für die Schichtung und im Bericht alternative Gruppierungen der Altersvariable vorgenommen wurden, sind hier auch drei unterschiedliche Tests für Alter angeführt. Es konnten weder für Bildung, Alter oder Geschlecht Unterschiede nachgewiesen werden. Lediglich für die Siedlungsdichte zeigt sich ein signifikanter Unterschied, denn anteilig mehr Personen aus gering besiedelten Gebieten sind ausgeschieden. Tatsächlich kommen fast 60 % aller 50 Retoursendungen aus Gebieten mit niedriger Siedlungsdichte, während für die inkludierten Adressen die Unterschiede zwischen den Häufigkeiten der drei Kategorien der Siedlungsdichte deutlich geringer ausfallen (niedrig: 31,5 %, mittel: 29,9 %, hoch: 38,6 %).
Der Ausschluss der 118 Adressen aus der Bruttostichprobe ist dementsprechend weitestgehend qualitätsneutral mit der Einschränkung, dass die Retouren für den ländlichen Raum überproportional hoch sind. Da mit 118 von 3750 aber nur ca. 3 % des ursprünglichen Adressenpools betroffen sind, ist nicht von einer unverhältnismäßigen Verzerrung auszugehen.
Über eine Analyse des Designgewichts kann das Design mit einer einfachen Zufallsstichprobe verglichen werden. Das Anpassungsgewicht für die zweite Piloterhebung berücksichtigt lediglich das Oversampling von bildungsfernen Schichten und sollte aufgrund des einfachen Designs gering ausfallen. Zur Prüfung wurden äquivalent zum ersten Bericht aus jedem Umfragemodul eine Frage ausgewählt, für welche der Designeffekt nach Lumley (2016) berechnet wurde. Tabelle 6 fasst die Ergebnisse zusammen. Der Designeffekt kann als Reduktionsfaktor der Stichprobengröße einer komplexen Stichprobe im Vergleich zu einer einfachen Zufallsstichprobe interpretiert werden. Im Kontext der Erhebung würde dies bedeuten:

$$n_{\mathit{eff}} = \frac{721}{1{,}14} = 632. \tag{1}$$

Mit anderen Worten, die über die PUMA mit dem Oversampling realisierte Nettostichprobe von 721 Personen entspricht einer effektiven Nettostichprobe von 632 Personen. Im Vergleich zu dem komplexen Stichprobendesign bei der ersten Piloterhebung ist der Designeffekt diesmal sehr gering. Der Designeffekt aus der ersten Erhebung betrug wenigstens 1,28, welcher die Nettostichprobe dieser Erhebung von 721 auf eine effektive Nettostichprobe von 563 reduzieren würde. Dementsprechend ist das einfachere Design dem komplexeren wenn möglich vorzuziehen.

Tabelle 7: Ausschöpfung mit und ohne Designgewichtungskorrektur

	Ausschöpfung	
	ungewichtet	gewichtet
Gruppe A	12,3 %	14,1 %
Gruppe B	21,6 %	24,6 %
Gruppe C	23,4 %	25,7 %
Gruppe D	29,9 %	32,5 %
Gruppe E	22,1 %	24,7 %
Gruppe F	15,1 %	16,0 %
Gruppe G	17,1 %	18,0 %

4.3 Non-Response-Fehler

Auch wenn im Abschnitt zum Stichprobenziehungsfehler Retoursendungen ausgeschlossen wurden, so muss zu Beginn dieses Abschnitts explizit darauf hingewiesen werden, dass deshalb nicht davon ausgegangen werden kann, dass die 3.632 Personen aus der bereinigten Bruttostichprobe alle erreicht worden sind. Es kann lediglich für die 771 Personen, welche den Fragebogen begonnen haben, mit Sicherheit gesagt werden, dass die Anschreiben einen Adressaten gefunden haben. Dass auch im Fall eines vollständig ausgefüllten Fragebogens nicht die richtige Person erreicht worden sein muss, lässt sich daran erkennen, dass nur für 82,5 % aller Personen der Nettostichprobe Geschlecht und Alter aus den Angaben im Fragebogen mit den Daten des ZMR übereinstimmen. Lässt man 1 Jahr Abweichung beim Alter zu, dann erhöht sich der Anteil auf 92,5 %, für die restlichen 7,5 % stimmt entweder das Geschlecht nicht oder das Alter weicht um mehr als 1 Jahr von den Daten des ZMR ab. Beides deutet eher darauf hin, dass die befragte Person nicht die über das Anschreiben adressierte Person ist.

Die Tabellen 7 und 8 weisen deutliche Unterschiede in der Ausschöpfung und den Abbrüchen über die Incentivegruppen aus. Die Gruppe A hat mit 12,3 % eine sehr geringe Ausschöpfung und fällt damit deutlich hinter alle anderen Gruppen mit Incentive zurück. Dies scheint ein starkes erstes Argument für Incentives zu sein. Eine zweite Beobachtung ist, dass die unkonditionalen Incentivegruppen nochmals deutlich besser abschneiden als die ausschließlich konditionalen. Die beste Gruppe nach der Ausschöpfung ist die Gruppe D, welche sowohl eine 2 EUR Münze als bedingungslosen Incentive im Folder und nach der Teilnahme an der Befragung noch die 5 EUR Münze zugesandt bekam. Ausgehend von der Ausschöpfung ist das Ergebnis scheinbar eindeutig, allerdings widersprechen die Unterschiede in den Abbrüchen dieser Schlussfolgerung. Bei den Abbrüchen sind es gerade die nach der Ausschöpfung schlechteren Gruppen A (kein Incentive), B (2 EUR Münze unkonditional) und G (10 EUR Gutschein konditional), welche anteilig die wenigsten Abbrüche verzeichnen. Aufgrund der sehr geringen Anzahl an Abbrüchen sind diese Ergebnisse aber höchstens als Tendenz interpretierbar.

Ausschöpfung allein ist nur im Fall einer Ausschöpfung von 100 % als alleiniges Qualitätsmerkmal geeignet, denn unabhängig von der Gesamtausschöpfung kann die Aus-

Tabelle 8: Ausschöpfung über bereinigte Stichprobe

Nettostichprobe	Incentivgruppe							Gesamt
	Gruppe A kein incentive	Gruppe B 2 Euro im Säckchen	Gruppe C 2 Euro in Folder	Gruppe D 2 Euro in Folder + 5 Euro im Säckchen konditional	Gruppe E 5 Euro im Säckchen	Gruppe F 5 Euro Münze im Säckchen konditional bei Beantwortung	Gruppe G 10 Eurogutschein konditional bei Beantwortung	
fehlend	416	338	363	329	294	323	798	2861
	86,8%	77,5%	74,4%	68,4%	76,4%	82,8%	82,0%	78,8%
gültig < 13 fehlend	59	94	114	144	85	59	166	721
	12,3%	21,6%	23,4%	29,9%	22,1%	15,1%	17,1%	19,9%
ungültig/Abbruch	4	4	11	8	6	8	9	50
	0,8%	0,9%	2,3%	1,7%	1,6%	2,1%	0,9%	1,4%
Gesamt	479	436	488	481	385	390	973	3632
	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

$\chi^2 = 72{,}912; df = 12; p < 0{,}001$

schöpfung über die Kategorien einzelner Variablen variieren und damit zu Verzerrungen führen. Tabelle 9 stellt die Ergebnisse einer designgewichteten logistischen Regression auf die Teilnahme an der Befragung dar, wobei alle Variablen dichotomisiert bzw. dummykodiert wurden, um die unstandardisierten Koeffizienten leichter vergleichbar zu machen. Junge Menschen konnten besser rekrutiert werden im Vergleich zu allen anderen Altersgruppen, wobei die 65- bis 74-Jährigen sich hier nochmals deutlich negativ von den anderen Altersgruppen abheben. Den stärkeren Effekt hat aber die Bildung, die eine ansteigende Teilnahmebereitschaft von niedrigen über mittleren hin zu hohem Bildungsniveau zeigt. Für Siedlungsdichte und Geschlecht sind keine Effekte nachweisbar. Personen die aktuell inaktiv sind auf dem Arbeitsmarkt (nicht erwerbstätig), konnten leichter rekrutiert werden, während dies für Arbeitslose besonders schwierig war. Außerdem sind AusländerInnen weniger vertreten.
Im Übergang von Modell 1 zu Modell 2 bleiben die Effekte über alle Variablen stabil und es bestätigt die Erkenntnis aus der Beschreibung der Ausschöpfung. Die Gruppe A und F unterscheiden sich nicht in der Ausschöpfung und Gruppe G zeigt nur einen schwachen Effekt im Vergleich zu allen anderen Effekten im Modell. Gruppe D erhöht die Teilnahmebereitschaft am stärksten, auch wenn der Unterschied zwischen Gruppe A und D nicht so deutlich ist wie zwischen niedriger und hoher Bildung.
Eine detailliertere Beschreibung der Zusammenhänge zwischen Rekrutierung und den soziodemographischen Variablen erfolgt über die Abbildungen 3a, 3b und 3c. Entlang der Grafiken zeigt sich, dass die Gruppe D zwar in allen drei Variablen mehrheitlich überdurchschnittliche Ausschöpfung erreicht, dass diese aber über die Ausprägungen der Variablen nicht gleichmäßig verteilt sind. Der kombinierte Incentive scheint weniger gut bei Personen mit geringer Bildung, den älteren Befragten und im ländlichen Raum zu funktionieren. Im Vergleich dazu sind die Verteilungen der Gruppe C erheblich harmonischer.
Für die Personen mit Hochschulabschluss wirken die Incentives gut, allerdings erreicht diese Gruppe in allen Incentivegruppen die durchschnittliche Ausschöpfung von knapp 20 %. Für die mittlere Bildungsgruppe sind besonders die bedingungslosen Incentives wirksam, während die Bildungsfernen in keiner Incentivegruppe die durchschnittlichen 20 % Ausschöpfung erreichen. Das Oversampling der Bildungsfernen war zweifelsfrei notwendig.
Bei den Altersgruppen ist auffallend, dass jede Altersgruppe in wenigstens einer Incentivegruppe die durchschnittlichen 20 % Ausschöpfung erreicht. Die Ausschöpfungen

Tabelle 9: Designgewichtete logistische Regressionskoeffizienten der Stichprobenausschöpfung von Brutto- zu Nettostichprobe (Standardfehler in Klammern).

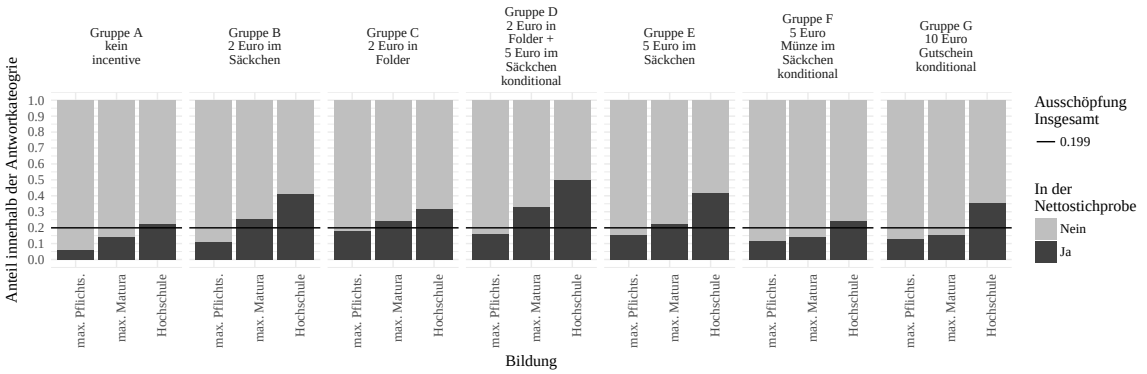
		Chance der Selektion	
		Model 1	Model 2
		logistic	*logistic*
	Konstante	−1,374***	−1,888***
		(0,044)	(0,051)
Geschlecht	Geschlecht	−0,001	−0,007
(Ref. = Männer)		(0,017)	(0,018)
Alter	25 bis 34 Jahre	−0,251***	−0,315***
(Ref. = 16 bis 24 Jahre)		(0,033)	(0,033)
	35 bis 44 Jahre	−0,379***	−0,426***
		(0,034)	(0,035)
	45 bis 54 Jahre	−0,242***	−0,280***
		(0,032)	(0,033)
	55 bis 64 Jahre	−0,280***	−0,320***
		(0,033)	(0,033)
	65 bis 74 Jahre	−0,739***	−0,790***
		(0,038)	(0,039)
Bildung	Lehre, berufsbildende Schule oder Matura	0,536***	0,560***
(Ref. = unbekannt oder maximal Pflichtschule)		(0,029)	(0,030)
	Hochschule	1,315***	1,355***
		(0,034)	(0,035)
Siedlungsdichte	mittel	−0,162***	−0,157***
(Ref. = niedrig)		(0,022)	(0,023)
	hoch	−0,133***	−0,141***
		(0,021)	(0,021)
Erwerbsstatus	Arbeitslos	−0,495***	−0,528***
(Ref. = erwerbstätig)		(0,052)	(0,053)
	Nicht erwerbstätig	0,059***	0,074***
		(0,023)	(0,023)
Staatsbürgerschaft	EU25 (ohne Ö)	−0,695***	−0,680***
(Ref. = Österreich)		(0,037)	(0,037)
	Nicht-EU	−1,098***	−1,159***
		(0,049)	(0,049)
Incentive	Gruppe B		0,732***
(Ref. = kein Incentive)			(0,037)
	Gruppe C		0,728***
			(0,036)
	Gruppe D		1,127***
			(0,035)
	Gruppe E		0,736***
			(0,038)
	Gruppe F		0,094**
			(0,040)
	Gruppe G		0,312***
			(0,033)
	Pseudo-R^2	0,041	0,062
	N	3.620	3.620
	Log Likelihood	−41.805,310	−40.915,920
	Akaike Inf. Crit.	83.640,610	81.873,840
Anmerkung:			*p<0,1; **p<0,05; ***p<0,01

Tabelle 10: Designgewichtete logistische Regressionskoeffizienten der Stichprobenausschöpfung von Brutto- zu Nettostichprobe nach Incentivegruppen (Standardfehler in Klammern).

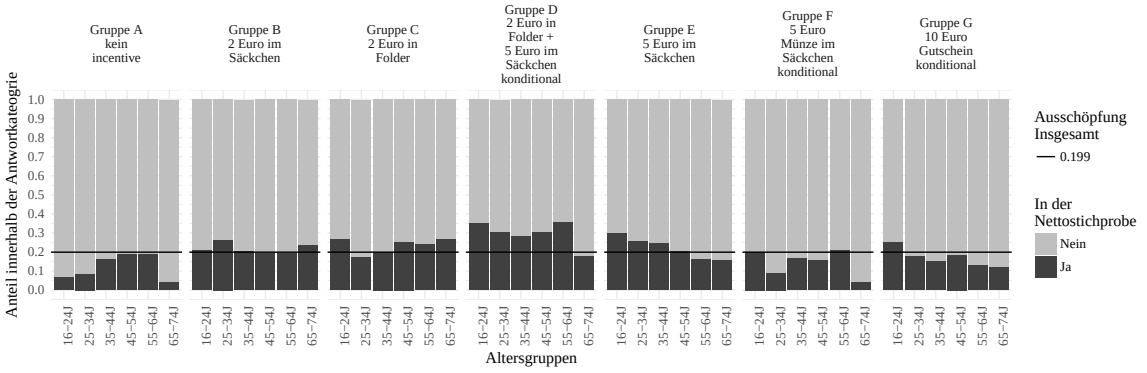
		Chance der Selektion						
		Gruppe A	Gruppe B	Gruppe C	Gruppe D	Gruppe E	Gruppe F	Gruppe G
		logistic	*logistic*	*logistic*	*logistic*	*logistic*	*logistic*	*logistic*
	Konstante	−3,088***	−1,912***	−0,730***	−1,393***	−0,539***	−0,890***	−1,407***
		(0,164)	(0,136)	(0,112)	(0,114)	(0,127)	(0,157)	(0,089)
Geschlecht (Ref. = Männer)	Geschlecht	0,294***	−0,236***	−0,041	−0,013	0,292***	−0,133**	−0,050
		(0,059)	(0,052)	(0,046)	(0,043)	(0,055)	(0,061)	(0,037)
Alter (Ref. = 16 bis 24 Jahre)	25 bis 34 Jahre	−0,164	0,332***	−0,509***	−0,451***	−0,735***	−0,931***	−0,221***
		(0,126)	(0,103)	(0,088)	(0,085)	(0,095)	(0,127)	(0,068)
	35 bis 44 Jahre	0,730***	0,003	−0,625***	−0,549***	−0,791***	−0,596***	−0,693***
		(0,118)	(0,109)	(0,089)	(0,086)	(0,106)	(0,114)	(0,073)
	45 bis 54 Jahre	1,000***	−0,144	−0,232***	−0,525***	−0,845***	−0,321***	−0,407***
		(0,112)	(0,108)	(0,082)	(0,085)	(0,101)	(0,105)	(0,067)
	55 bis 64 Jahre	1,079***	−0,425***	−0,323***	−0,398***	−1,275***	0,319***	−0,894***
		(0,113)	(0,104)	(0,083)	(0,087)	(0,103)	(0,105)	(0,073)
	65 bis 74 Jahre	−0,327**	−0,625***	−0,596***	−1,202***	−0,867***	−0,868***	−1,069***
		(0,155)	(0,112)	(0,091)	(0,100)	(0,110)	(0,156)	(0,080)
Bildung (Ref. = unbekannt oder maximal Pflichtschule)	Lehre, berufsbildende Schule oder Matura	0,945***	0,938***	0,382***	0,884***	0,324***	0,282***	0,389***
		(0,108)	(0,086)	(0,073)	(0,076)	(0,083)	(0,104)	(0,061)
	Hochschule	0,928***	1,839***	0,777***	1,758***	1,091***	1,013***	1,568***
		(0,122)	(0,109)	(0,088)	(0,089)	(0,099)	(0,117)	(0,071)
Siedlungsdichte (Ref. = niedrig)	mittel	−0,437***	0,035	−0,246***	0,543***	−0,640***	−0,467***	−0,102**
		(0,073)	(0,068)	(0,056)	(0,060)	(0,072)	(0,077)	(0,047)
	hoch	−0,749***	0,354***	−0,542***	0,572***	−0,287***	−0,555***	−0,090**
		(0,073)	(0,062)	(0,055)	(0,054)	(0,062)	(0,072)	(0,045)
Erwerbsstatus (Ref. = erwerbstätig)	Arbeitslos	0,188	−0,296*	−0,744***	−1,210***	−1,306***	−0,715***	0,144
		(0,169)	(0,160)	(0,165)	(0,146)	(0,178)	(0,174)	(0,090)
	Nicht erwerbstätig	−0,304***	0,947***	0,277***	−0,014	−0,478***	−0,830***	0,323***
		(0,073)	(0,069)	(0,054)	(0,056)	(0,077)	(0,084)	(0,051)
Staatsbürgerschaft (Ref. = Österreich)	EU25 (ohne Ö)	−0,133	−1,772***	−0,455***	−0,567***	−1,189***	−0,707***	−0,628***
		(0,106)	(0,161)	(0,091)	(0,090)	(0,104)	(0,114)	(0,081)
	Nicht-EU	−0,313**	−16,412	−3,098***	−0,540***	−1,922***	−0,228	−0,958***
		(0,141)	(163,338)	(0,294)	(0,091)	(0,158)	(0,168)	(0,091)
	Pseudo-R^2	0,086	0,108	0,055	0,071	0,098	0,081	0,055
	N	478	434	485	481	383	389	970
	Log Likelihood	−4.088,402	−4.722,369	−6.080,852	−6.562,465	−4.500,955	−3.731,494	−9.871,353
	Akaike Inf. Crit.	8.206,805	9.474,737	12.191,700	13.154,930	9.031,911	7.492,988	19.772,710

Anmerkung: *p<0,1; **p<0,05; ***p<0,01

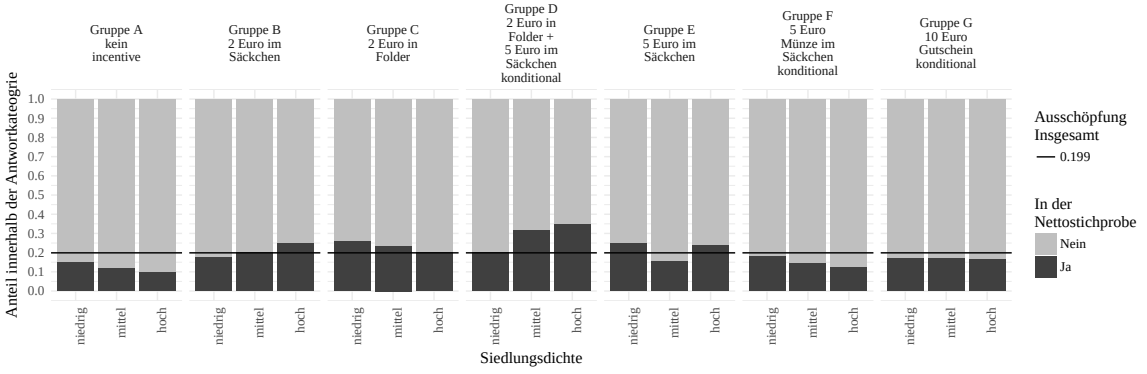
Abbildung 3: Balkendiagramm zur Ausschöpfung über Antwortkategorien nach Incentivegruppen



(a)



(b)



(c)

Tabelle 11: R-Indikator und Pseudo-R^2 der designgewichteten logistischen Regression für Ausschöpfung aus Brutto- zu Nettostichprobe nach Incentivegruppe

Gruppe	n	Ausschöpfung (gewichtet)	R-Indikator	Vertrauens-intervall	Maximaler Bias	Maximaler Kontrast	Pseudo-R^2
A	479	14,1 %	0,865	(0,788-0,941)	0,479	0,558	0,081
B	436	24,6 %	0,760	(0,678-0,843)	0,486	0,645	0,129
C	488	25,5 %	0,862	(0,783-0,941)	0,271	0,363	0,053
D	481	32,5 %	0,776	(0,694-0,859)	0,344	0,510	0,096
E	385	24,9 %	0,769	(0,678-0,861)	0,463	0,617	0,111
F	390	16,0 %	0,855	(0,761-0,949)	0,453	0,539	0,066
G	973	18,0 %	0,846	(0,788-0,904)	0,427	0,521	0,067

über die Altersgruppen sind mehrheitlich tendenziell homogen, wobei in Gruppe D mit Ausnahme für die älteste Personengruppe die höchsten Ausschöpfungen in allen Altersgruppen erzielt wurden. Danach zeigen die unkonditionalen Incentives bessere Ergebnisse als die konditionalen Incentives und Gruppe A ohne Incentive. Die Situation für die Siedlungsdichte ist ähnlich, wobei hier auffällt, dass die Personen in dicht besiedelten Gebieten stärker auf die Incentives reagieren, d.h. die Streuung dieser Gruppe über die Incentivegruppen ist höher als die Streuung der Gruppe im ländlichen Raum.
Die Ergebnisse der grafischen Analyse werden durch den im R-Indikator bestätigt. Beim R-Indikator zeigt die Gruppe B mit 0,760 den geringsten Wert für Repräsentativität. Dies lässt sich aus den stärken Schwankungen zwischen den Ausschöpfung über die Kategorie der Variablen erklären, welche der R-Indikator berücksichtigt und sich auch im Pseudo-R^2 zeigt. Interessanterweise ist die Gruppe C die mit Abstand repräsentativste Incentivegruppe über alle Kenngrößen hinweg. Der R-indikator allein ist genauso groß wie für Gruppe A, aber im Vergleich von maximalen Bias, maximalem Kontrast und Pseudo-R^2 manifestiert sich die höhere Repräsentativität von Gruppe C.
Die Non-Response-Analyse hat verdeutlicht, dass es neben der Ausschöpfung einer detaillierteren Betrachtung bedarf, um die Qualität der Erhebung zu beurteilen. Insgesamt kann festgehalten werden, dass die Gruppe C, welche im Vergleich der Ausschöpfung noch unscheinbar im Mittelfeld lag, die höchste Repräsentativität aufweist, denn über die berücksichtigten soziodemographischen Variablen variiert die Ausschöpfung am wenigsten.

4.4 Anpassungsfehler

Eine Möglichkeit zur Betrachtung des Anpassungsfehlers sind Gewichtungen. Hier können zwei Arten der Gewichtungen betrachtet werden. Einerseits ermöglicht die Designgewichtung eine Betrachtung der potentiellen Ausschöpfung ohne Oversampling von Personen mit niedriger Bildung (vgl. Tabelle 7). Unter Berücksichtigung der Designgewichtung steigt die Ausschöpfung in allen Incentivegruppen leicht an. Dies ergibt sich daraus, dass mit dem Oversampling mehr Personen mit niedriger Bildung adressiert wurden und diese die erwartete geringere Ausschöpfung aufwei-

Tabelle 12: Vergleich des Variations- undQuartilsdispersionskoeffizienten

Erhebung	Gruppe	Incentive unkonditional	Incentive konditional	Variations-koeffizient	Quartils-dispersions-koeffizient
Pilot I		Broschüre		0,932	0,599
Pilot I		2 Euro im Säckchen		0,732	0,525
Pilot I		5 Euro im Säckchen		0,990	0,298
Pilot I		10 Euro Gutschein		0,597	0,642
Pilot II	A			1,002	1,173
Pilot II	B	2 Euro im Säckchen		0,787	0,759
Pilot II	C	2 Euro in Folder		0,363	0,444
Pilot II	D	2 Euro in Folder	5 Euro im Säckchen	0,643	0,505
Pilot II	E	5 Euro im Säckchen		0,762	0,600
Pilot II	F		5 Euro im Säckchen	0,617	0,708
Pilot II	G		10 Euro Gutschein	0,401	0,535

sen. Außerdem ermöglicht erst die gewichtete Ausschöpfung den Vergleich zu anderen Erhebungen.
Anderseits können die inversen Auswahlwahrscheinlichkeiten aus den Regressionsmodellen verwendet werden, denn in den Regressionsmodellen sind die Mehrzahl der Variableninkludiert, welche bei der Bestimmung der Hochrechnungsgewichte genutzt werden. Da die Hochrechnungsgewichte aber über die gesamte Nettostichprobe kalibriert wurden, ist diese Gewichtung für den Vergleich der Gruppen unbrauchbar. Als Näherungswert können die Regressionsmodelle auf die Variablen Geschlecht, Alter, Siedlungsdichte, Bildung und Erwerbsstatus gekürzt werden. Diese Modelle wurden für jede Incentivegruppe separat getestet und im Anschluss wurden die inversen Auswahlwahrscheinlichkeiten für die Fälle der Nettostichprobe als Grundlage zur Berechnung des Variations- und Quartilsdispersionskoeffizienten herangezogen.
Sowohl Variations- als auch Quartilsdispersionskoeffizienten sind relative Streuungsmaße, die den Vergleich über unterschiedliche Stichproben ermöglichen. Der Variationskoeffizient bezieht sich auf den Mittelwert und ist definiert als

$$v = \frac{s}{\bar{x}}, \tag{2}$$

mit s für die Standardabweichung und $\bar{x}$ für den Mittelwert (Benesch, 2012, S. 48). Der Variationskoeffizient ist nicht robust gegen Ausreißer und deshalb soll an dieser Stelle auch noch der Quartilsdispersionskoeffizient eingeführt werden. Der Quartilsdispersionskoeffizient (QDK) berechnet sich aus dem Verhältnis der Quartile (Benesch, 2012, S. 47).:

$$QDK = \frac{x_{0,75} - x_{0,25}}{x_{0,5}}. \tag{3}$$

Die Ergebnisse aus Tabelle 12 bestätigen den Eindruck der grafischen Interpretation mit Bezug auf die Gruppe C aus der zweiten Piloterhebung. Der niedrige Variationskoeffizient verweist daraf, dass die unterschiedlichen Subpopulationen eine ähnliche Chance haben rekrutiert zu werden. Damit sinkt auch die Notwendigkeit von

Tabelle 13: Notwendige Bruttostichproben basierend auf der Ausschöpfung

	Effektive Stichprobe	Design-effekt	Notw. Stichprobe mit Designeffekt	Ausschöpfung	Notw. Brutto-stichprobe
Gruppe A	1.000	1,14	1.145	12,3 %	9.294
Gruppe B	1.000	1,14	1.145	21,6 %	5.310
Gruppe C	1.000	1,14	1.145	23,4 %	4.900
Gruppe D	1.000	1,14	1.145	29,9 %	3.824
Gruppe E	1.000	1,14	1.145	22,1 %	5.185
Gruppe F	1.000	1,14	1.145	15,1 %	7.567
Gruppe G	1.000	1,14	1.145	17,1 %	6.710

Anpassungsgewichten zur nachträglichen Korrektur. Auch im Vergleich mit dem robusten QDK ist die Gruppe C überdurchschnittlich gut. Die Unterschiede zwischen den restlichen Gruppen ist weniger deutlich ausgeprägt. Lediglich Gruppe A würde eine deutlich stärkere Anpassung benötigen. Die Bedeutung eines geringen Variationskoeffizienten kann in diesem Kontext nur nochmals hervorgehoben werden, denn die Korrektur über Gewichte ist ein wichtiges Mittel zur Verbesserung der Datenqualität, aber diese Methode kann die Bemühungen zur Erhöhung der Ausschöpfung und der Repräsentativtät der Rohdaten nicht kompensieren.

4.5 Kosten

Abschließend werden noch die Kosten auf Basis der gewonnen Erkenntnisse geschätzt. Auch an dieser Stelle wird äquivalent zum ersten Bericht zunächst eine effektive Stichprobe von 1.000 Personen und durch die globale Umlegung des Designeffekts von 1,14 eine realisierte Nettostichprobe von 1.145 für alle Incentivegruppen als Annahme eingeführt. Im Anschluss wird über die beobachtete Ausschöpfung die notwendige Bruttostichprobe ermittelt (vgl. Tabelle 13). Und ausgehend von dieser notwendigen Bruttostichprobe werden in Abhängigkeit von der Art des Incentives die jeweiligen Gesamtkosten mit Blick auf eine mögliche Wiederholungsbefragung 2017 grob geschätzt. Die Stückkosten variieren in den Portokosten in Abhängigkeit vom Gewicht der Sendung, im Inhalt durch den Wert des Incentives und die Herstellungskosten. Insgesamt wird deshalb nur zwischen groben Planungsstückkosten für die Netto- und Bruttostichprobe unterschieden, welche etwas höher ausfallen als in der ersten Pilot Erhebung.[3]
Unkonditionale Incentives von Gruppe E sind deutlich mit den höchsten Kosten verbunden, während die Gruppe A trotz der deutlich größeren notwendigen Bruttostichprobe die geringsten Kosten verursacht. Die Gruppe F und G sind mit knapp 10 % höheren Kosten verbunden als Gruppe A. Auf einem ähnlichen Kostenniveau liegen

[3]Die Kostenspezifikation sind gerundete Planungsgrößen der Statistik Austria zum Stand der Berichtlegung. Die tatsächlichen Kosten z.B. der Produktion von Foldern hängen teilweise auch von den jeweiligen Stückzahlen ab. Die tatsächlichen Stückkosten für die zweite PUMA Erhebung lagen deutlich höher aufgrund der geringen Stückzahlen und Einmalkosten für Layout etc. Portogebüren wurden im Jahr 2017 empfindlich angehoben usw. Die hier angesetzten Planungsgrößen dienen lediglich zur Illustration von Größenordnungen potentieller Kosteneffekte.

Tabelle 14: Kostenschätzung basierend auf den notwendigen Bruttostichproben

	Bruttostichprobe			Nettostichprobe			
	Notwendig	Stückkosten	Samplekosten	Realisiert	Stückkosten	Samplekosten	Gesamtkosten
Gruppe A	9.294	2,60 EUR	24.164 EUR	1.145	0,00 EUR	0 EUR	24.164 EUR
Gruppe B	5.310	6,90 EUR	36.637 EUR	1.145	0,00 EUR	0 EUR	36.637 EUR
Gruppe C	4.900	7,90 EUR	38.712 EUR	1.145	0,00 EUR	0 EUR	38.712 EUR
Gruppe D	3.824	7,90 EUR	30.208 EUR	1.145	7,50 EUR	8.588 EUR	38.795 EUR
Gruppe E	5.185	9,90 EUR	51.332 EUR	1.145	7,50 EUR	8.588 EUR	59.919 EUR
Gruppe F	7.567	2,60 EUR	19.674 EUR	1.145	7,00 EUR	8.015 EUR	27.689 EUR
Gruppe G	6.710	2,60 EUR	17.446 EUR	1.145	11,10 EUR	12.710 EUR	30.155 EUR

Anm.: Laut Statistik Austria lagen die tatsächlichen Kosten vor allem für den Folder aufgrund der geringen Auflage und einmaliger Kosten für das Layout deutlich höher, sind aber für folgende Erhebungen wesentlich geringer anzusetzen. Für den Folder für Gruppe C und D sollten die Kosten bei einer wiederholten Anwendung nur ca. ein Drittel betragen.

Tabelle 15: Vergleich Pilot I und Pilot II Gesamtstichprobe

	R-Indikator	max. Design Effekt	min. Design Effekt	Ausschöpfung
PILOT I (Q2-2016)	0,73	1,53	1,28	0,26
PILOT II (Q4-2016)	0,77	1,19	1,14	0,22

die Gruppen B bis D, allerdings sind diese Varianten um ca. 10.000 EUR teurer als Gruppe A. An dieser Stelle ist auch darauf hinzuweisen, dass Gruppe B und Gruppe C sich nur durch die Verpackung nicht aber durch die Höhe des Incentives unterschieden haben. Dementsprechend kann ein qualitativ hochwertigeres Anschreiben durchaus die Qualität erhöhen. Damit ist im Besonderen der Folder eine sehr attraktive Ergänzung der Rekrutierungsbemühungen.

5 Vergleich Pilot I und Pilot II

Ein zentrales Ziel der Piloterhebungen ist die Identifikation eines qualitativ hochwertigen Erhebungsmodus in Abwägung gegen Praktikabilität und Kosten. Dabei lässt sich im Vergleich von den beiden Piloterhebung feststellen, dass die Ziehung aus dem ZMR qualitativ bessere Ergebnisse erzielt, wie sich in R-Indikator und Designeffekt deutlich erkennen lässt (vgl. Tabelle 15). Gleichzeitig war die Ausschöpfung in der ersten Piloterhebung höher, was sich letztlich in den sehr großen Differenzen der geschätzten Kosten niederschlägt. Außerdem besitzt das Design der Pilot I Erhebung den Vorteil, dass die Stückkosten für die Rekrutierung über das Telefon mit 2 EUR sehr gering sind und Porto und Incentivekosten erst nach einer Vorabrekrutierung anfallen. Bei der Auswahl über das ZMR werden die unkonditionalen Incentives an alle Personen gesendet und damit sind die Kosten deutlich höher. Während Pilot I bei Ausschöpfung und Kosten bessere Ergebnisse liefert, können die Pilot II Erhebungen als qualitativ hochwertiger betrachtet werden mit Bezug auf höhere Repräsentativität und geringerem Designeffekt.
Die Diskussion im Rahmen des Anpassungsfehlers ist ein weiteres Indiz für die besseren Eigenschaften der Stichproben aus dem ZMR. Wenn die Randverteilungen der amtlichen Statistik die Referenz für die Hochrechnungsgewichte sind und die Grund-

Tabelle 16: Vergleich Pilot I und Pilot II für bedingungslosen 5 EUR Münzen Incentive

	R-Indikator	max. Design Effekt	min. Design Effekt	Ausschöpfung	Kosten
PILOT I (Q2-2016)	0,70	1,47	1,32	0,30	29.208 EUR
PILOT II (Q4-2016)	0,77	1,43	1,03	0,25	59.919 EUR

Anm.: In Pilot I waren die Planungswerte für den Versand (inkl. Porto und Druckkosten) geringer. Um einen sinnvollen Vergleich zu ermöglichen, sind die Incentivestückkosten für Pilot I mit denen aus der Pilot II gleichgesetzt worden.

Tabelle 17: Vergleich Pilot I und Pilot II für bedingungslosen 2 EUR Münzen Incentive

	R-Indikator	max. Design Effekt	min. Design Effekt	Ausschöpfung	Kosten
PILOT I (Q2-2016)	0,73	1,53	1,12	0,28	23.531 EUR
PILOT II (Q4-2016)	0,76	1,19	1,12	0,25	36.637 EUR

Anm.: In Pilot I waren die Planungswerte für den Versand (inkl. Porto und Druckkosten) geringer. Um einen sinnvollen Vergleich zu ermöglichen, sind die Incentivestückkosten für Pilot I mit denen aus der Pilot II gleichgesetzt worden.

gesamtheit über diesen Referenzpunkt mit hoher Präzision abgebildet werden kann, dann darf ein geringes Anpassungsgewicht als Qualitätskriterium einer Stichprobe gelten und sollte Vorrang vor Ausschöpfung und Kosten besitzen. Wichtig bei dieser Interpretation ist aber, dass der Umkehrschluss nicht theoretisch ableitbar ist. Ein größeres Anpassungsgewicht muss nicht mit einer geringeren Repräsentativität verbunden sein. Dieses Spannungsverhältnis ist für die Betrachtung der Kosten und der nachhaltigen Qualität der Datenerhebung zu berücksichtigen.

Im Kontext des Vergleichs von Pilot I und Pilot II Erhebung ist davon auszugehen, dass die statistischen Kennzahlen Evidenz für die Datenerhebung über das ZMR liefern, aber der für die Forschung relevanteste Fehler, die Verzerrung in der Streuung der erhobenen Variablen, die über die aus der amtlichen Statistik bekannten Randverteilungen hinausgehen, kann hiermit nicht abschließend erörtert werden. An dieser Stelle soll abschließend nur darauf verwiesen werden, dass der Vergleich zwischen den beiden Designs von Pilot I und Pilot II an verschieden Stellen nur begrenzt möglich ist, da zu viele Elemente der Designs variieren.[4]

Der Vergleich der Incentivegruppen ist nur eingeschränkt möglich, da nur zwei Gruppen identische Incentives angeboten bekamen. Die Rekrutierung selbst ist ein wesentlicher Designunterschied und kann hier durch die starken Unterschiede der Incentivegruppen nur abgeschätzt werden. Bei der Piloterhebung I erfolgte der Kontakt über eine schon bestehende Umfrage mit verpflichtender Teilnahme und deshalb ist einerseits fraglich, ob die Befragten den Unterschied zwischen den Befragungen wirklich realisierten und anderseits, wie groß der Rekrutierungseffekt durch die Befragungsermüdung oder die Reputation der vorherigen Erhebung war.

Bei der Kontaktaufnahme der Pilot II Studie über das ZMR ist davon auszugehen, dass ein Großteil der Adressen, welche aus dem ZMR abgerufen wurden, auch erreicht wurden. Allerdings deuten die Unterschiede in den Merkmalen Alter und Geschlecht

[4]Bspw. sind die Befragungen im Modus zwar identisch aber nicht im Inhalt. Dies ist besonders deshalb relevant, da der Feldbericht der Statistik Austria (Till, 2017) aufzeigt, dass in der Pilot II Erhebung ein Fragemodul mit Vignetten zum Einsatz kam, welches zu Irritationen bei den Befragten führte. Auch wenn aus Tabelle 8 ersichtlich ist, dass die Anzahl der Abbrüche während der Pilot II Erhebung vernachlässigbar gering ist.

darauf hin, dass die befragte und die adressierte Person nicht immer identisch sind.[5] Außerdem ist vollkommen unbekannt, wie viele Anschreiben aus der bereinigten Bruttostichprobe von 3.632 von der adressierten Person wahrgenommen wurden. Während in der Pilot I Studie ein direkter telefonischer Kontakt zur Rekrutierung bestand, ist dies für die Pilot II Erhebung nicht garantiert. Dementsprechend wurde im Bericht auch die einfachste Ausschöpfung nach der Empfehlung der AAPOR (2016) gewählt.

6 Bereitschaft zur wiederholten Teilnahme

Die Bereitschaft zur wiederholten Teilnahme an einer Befragung wurde in der zweiten Pilot Erhebung über die Weitergabe der E-Mailadresse erfasst. Die Befragten wurden gebeten am Ende der Befragung mit Verweis auf zukünftige Befragungen eine E-Mailadresse anzugeben, falls sie bereit sind daran teilzunehmen. Die Angabe der E-Mailadresse soll dementsprechend an dieser Stelle als Bereitschaft zur wiederholten Teilnahme operationalisiert werden. In Tabelle 18 wurde ein vergleichbares logistisches Regressionsmodell für die Teilnahmebereitschaft wie vorher für die Selektivität gerechnet.
Die Altersgruppen unterscheiden sich alle gegen die Referenzgruppe der jüngsten Befragten, aber Unterschiede unter den restlichen Altersgruppen sind nicht nachweisbar. Bei der Bildung zeigt sich einmal mehr ein direkt proportionaler Zusammenhang, wo die Bildungsfernen die geringste Bereitschaft zur wiederholten Teilnahme zeigen und die Personen mit Hochschulabschluss die stärkste. Die Siedlungsdichte zeigt nur einen schwachen Zusammenhang auf, welcher eine geringere Bereitschaft bei der urbanen Bevölkerung nahelegt. Der Erwerbsstatus ist unabhängig von der Wiederholungsbereitschaft. Interessanter ist die geringere Bereitschaft von Personen ohne österreichische Staatsbürgerschaft wiederholt teilzunehmen. Gemeinsame mit den Rekrutierungsschwierigkeiten bei EU25 Staatsangehörigen kann dies eine Herausforderung für ein potentielles Onlinepanel werden. Die Effekte bleiben ähnlich wie bei den Modellen zur Selektivität über die Einführung der Dummyvariablen für die Incentivegruppen stabil.
Bei den Incentivegruppen zeigen die Gruppen mit konditionalem Incentive bessere Rekrutierungschancen für die Wiederholungsbefragung (Gruppe D, F und G). Überraschenderweise ist die Chance für die wiederholte Teilnahme bei der Incentivegruppe E mit der unkonditionalen 5 EUR Münze ebenfalls höher als die anderen unkonditionalen Gruppen B und C oder die incentivefreie Gruppe A. Dieses Ergebnis steht durchaus im Einklang mit der Literatur zur Incentivepraxis in Panelstudien (Singer & Ye, 2013; Laurie & Lynn, 2009; Göritz, 2006), die den Einsatz von Incentives während der Panelstudie variieren und oft Mischformen von unkonditionalen und konditionalen Incentives einsetzen. Unkonditionale Incentive haben sich hier als besonders

[5] Dieses Problem ist durchaus korrigierbar, denn die befragte Person ist sehr wahrscheinlich Teil der Grundgesamtheit und die Schichtung erfolgte über Bildung, Alter und Siedlungsdichte. Die Siedlungsdichte ist für adressierte und befragte Person identisch, während Alter und Bildung divergieren können. Da diese Informationen aber erhoben werden, kann dies nachträglich korrigiert werden. Mit einem Anteil von 7,6 % der Stichprobe kann dies ungewollte Effekte auf die ursprüngliche Schichtung haben.

Tabelle 18: Gewichtete logistische Regressionskoeffizienten der Bereitschaft zur Teilnahme an einer Wiederholungsbefragung (Standardfehler in Klammern).

		Chance der Selektion	
		Model 1	Model 2
		logistic	*logistic*
	Konstante	−1,882***	−2,238***
		(0,057)	(0,067)
Geschlecht (Ref = Männer)	Geschlecht	0,075***	0,056**
		(0,024)	(0,024)
Alter (Ref = 16 bis 24 Jahre)	25 bis 34 Jahre	−0,668***	−0,702***
		(0,041)	(0,041)
	35 bis 44 Jahre	−0,976***	−1,004***
		(0,044)	(0,044)
	45 bis 54 Jahre	−0,770***	−0,811***
		(0,040)	(0,041)
	55 bis 64 Jahre	−1,011***	−1,043***
		(0,043)	(0,043)
	65 bis 74 Jahre	−1,129***	−1,187***
		(0,050)	(0,050)
Bildung (Ref = unbekannt oder maximal Pflichtschule)	Lehre, berufsbildende Schule oder Matura	0,606***	0,612***
		(0,040)	(0,041)
	Hochschule	1,228***	1,226***
		(0,047)	(0,047)
Siedlungsdichte (Ref = niedrig)	mittel	−0,161***	−0,162***
		(0,029)	(0,029)
	hoch	−0,434***	−0,463***
		(0,029)	(0,029)
Erwerbsstatus (Ref = erwerbstätig)	Arbeitslos	−0,048	−0,076
		(0,062)	(0,062)
	Nicht erwerbstätig	−0,032	−0,008
		(0,031)	(0,031)
Staatsbürgerschaft (Ref = Österreich)	EU25 (ohne Ö)	−0,645***	−0,642***
		(0,051)	(0,051)
	Nicht-EU	−1,049***	−1,106***
		(0,066)	(0,067)
Incentive (Ref = kein Incentive)	Gruppe B		0,182***
			(0,054)
	Gruppe C		0,258***
			(0,051)
	Gruppe D		0,903***
			(0,047)
	Gruppe E		0,450***
			(0,052)
	Gruppe F		0,346***
			(0,052)
	Gruppe G		0,487***
			(0,044)
	Pseudo-R^2	0,040	0,049
	N	3.620	3.620
	Log Likelihood	−26.040,420	−25.791,460
	Akaike Inf. Crit.	52.110,840	51.624,920
Anmerkung:			*p<0,1; **p<0,05; ***p<0,01

effizient für die Erstrekrutierung herausgestellt, aber eine ideale Strategie ist bisher nicht identifizierbar (Laurie & Lynn, 2009, S. 230).

7 Schlussfolgerungen und Empfehlungen

Insgesamt unterscheiden sich die Gruppen teils deutlich in der Qualität und gleichzeitig sind die Ergebnisse für die unterschiedlichen Teile der Evaluation sehr unterschiedlich. Dementsprechend geht einer Reihung der Incentivegruppen die Anordnung der Evaluationskriterien voraus. Während der Diskussion wurde deutlich, dass nicht alle Probleme statistisch analysierbar und das einige Fehlerkomponenten nicht kontrollierbar sind. Der Vergleich zu hochwertigen alternativen Statistiken, wie der Vergleich der Hochrechnungsgewichte mit Bezug auf die amtliche Statistik, erschöpft sich ebenfalls im Vergleich bekannter Merkmale.
Der Bericht stützt sich wesentlich auf die Betrachtung des Non-Response-Fehlers über die Ausschöpfung und den R-Indikator, den Anpassungsfehler als Indikator der Verzerrung über die Analyse von Designeffekt und Hochrechnungsgewichten und auf eine Kostenschätzung. Idealerweise würde eine Gruppe in allen Bereichen deutlich bessere Resultate liefern als die Vergleichsgruppen. Durch die heterogenen Teilergebnisse muss ein Kriterium für die Rangfolge definiert werden. Dies soll zunächst die möglichst geringste Verzerrung für eine Querschnittsdatenerhebung sein. Die im letzten Abschnitt dargelegten Überlegungen zur Wiederholungsbefragung verdeutlichen bereits die Grenzen dieser Argumentation für die Zielstellung der Rekrutierungsexperimente.
Unter Berücksichtigung der Ergebnisse des R-Indikators und auch der graphischen Interpretation lässt sich die Gruppe C mit dem unkonditionellen 2 EUR Folderincentive als beste Option bezeichnen. Während sich die anderen Gruppen bezüglich R-Indikator, maximalem Bias und maximalen Kontrast weniger deutlich unterscheiden, wird an diesen Kenngrößen die Qualität der Gruppe C deutlich. Ein wichtiges Ergebnis ist auch, dass Gruppe D die mit Abstand höchste Ausschöpfung nur auf Kosten einer Verzerrung der Verteilung der soziodemographischen Merkmale realisiert werden kann.
Außerdem bestätigt sich das Ergebnis aus der ersten Pilotstudie, dass Bildung einen wesentlichen Einfluss auf die Teilnahmebereitschaft und auch die Bereitschaft zur wiederholten Teilnahme hat. Der Bildungseffekt war prominent in allen Regressionsmodellen und hat noch größeren Einfluss auf die Rekrutierung als die unterschiedlichen Incentives. Dementsprechend ist das durch Statistik Austria vorgenommene Oversampling sehr zu begrüßen, da damit die Streuung der finalen Gewichte verringert wird und das Oversampling dazu beiträgt, dass der Designeffekt II deutlich geringer ist als in der ersten Piloterhebung.
Beim Alter unterscheiden sich die jüngste und die älteste Altersgruppe von den mittleren. Die jungen Befragten reagieren durchgängig positiv auf Incentives und sind ohne Incentive nur schwer zu rekrutieren. Die ältesten Befragten sind ebenfalls ohne Incentive kaum rekrutierbar. Die unkonditionalen Incentives wirken und gleichzeitig zeigt sich, dass die konditionalen Incentives mit einer geringeren Bereitschaft zur Teilnahme einhergehen.

Das schlechte Abschneiden der konditionalen Incentives macht deren Anwendung für die Erstrekrutierung eigentlich obsolet, denn die incentivelose Variante ist nur unwesentlich schlechter aber gleichzeitig wesentlich preiswerter. Eine incentivelose Rekrutierung scheint ebenfalls nicht zielführend, da damit sowohl junge als alte Menschen und bildungsferne Personen sehr wahrscheinlich unterrepräsentiert sind. Die Ausschöpfungen sind dabei deutlich unterhalb der 10 %. Zwar konnten Teile dieser Verzerrung durch das Oversampling aufgefangen werden, aber durch die deutlich höhere Ausschöpfung bei nur geringfügig schlechterer Repräsentativität verweist Gruppe D die incentivelose Variante auf Platz 3.
Die Tatsache, dass mit der Gruppe C eine so ausgezeichnete Stichprobe gewonnen werden konnte, zeigt auch, dass auf die ethisch höchst problematischen differenzierten Incentives verzichtet werden kann. Eine zentrale Herausforderung für zukünftige Erhebungen ist die Wiederholung dieser Ergebnisse für die Gruppe C.
Für die Rekrutierungen zu Querschnittsdaten kann Gruppe C somit eindeutig empfohlen werden. Allerdings sollten die Ergebnisse auch mit der Perspektive auf eine Panelbefragung betrachtet werden. Hier zeigt sich bei der Wiederholungsbereitschaft, dass die zweitplatzierte Gruppe D, wie aus der Literatur bekannt, deutlich besser abschneidet. Singer und Ye (2013) belegen, dass ein höherer Incentive bei der Erstrekrutierung in Panels mittel- und langfristig zu höheren Ausschöpfungen führen können. Dementsprechend können die hohen Kosten für den Folder in Kombination mit einem konditionalen Incentive auch als anfängliche Investition betrachtet werden, welche die Motivation der Befragten nachhaltig verbessert. Der Geldwert des Incentives ist nur ein Aspekt zur längerfristigen Bindung in Panelstudien und alternative Mechanismen, wie etwa die Überzeugung in die Relevanz und Seriösität des Surveys, zeigen ebenfalls einen signifikanten Einfluss (Laurie & Lynn, 2009).
Eine weitere Möglichkeit zur Steigerung der Ausschöpfung und Qualität der Daten, der im Rahmen des Experiments nur universell über die Erinnerungsschreiben umgesetzt wurde, ist die Nachverfolgung von Non-Response. Da über das ZMR die Adressen der Befragten bekannt sind, kann die Nachverfolgung der Befragten noch über ein Erinnerungsschreiben hinausgehen. Im Kontext der Erfahrung aus den Piloterhebungen bieten sich grundsätzlich unterschiedliche Möglichkeiten an. Einerseits ist der Einsatz alternativer Befragungsmodi eine Option. Messer und Dillman (2011) testen unterschiedliche Kombinationen von Befragungsmodi in einem sogenannten web-plus-mail-Design (Internet und selbstadministrierter per Post übermittelter Fragebogen) und Nachverfolgung und erreichen über postalische Befragung bis zu ca. 70 %, wobei die Steigerungen über die alternativen Befragungsmodi zwischen ca. 15 % und 20 % liegen (vgl. Tabelle 2, Messer & Dillman, 2011, S. 437). Mit alternativen Befragungsmodi könnten nicht nur Personen rekrutiert werden, welche überhaupt nicht an der Befragung teilgenommen haben, sondern auch Personen, welche die Befragung aufgrund des Gerätes abgebrochen haben.
Eine zweite Möglichkeit der Nachverfolgung ist die Investition in mehr Kontaktversuche. Die web-plus-mail-Designs Messer und Dillman (2011) zeigen hier, dass gerade die Probleme mit den Abbrüchen bei Onlinesurveys aufgrund des Gerätetyps kompensiert werden können. Anstatt die Bruttostichprobe zu vergrößern, um eine gewünschte Nettostichprobe zu realisieren, könnte auch in die Ausschöpfung direkt investiert werden. Mögliche Strategien umfassen den bereits angesprochenen alternativen Befra-

gungsmodus, aber auch Adressnachverfolgung, mehrfache Anschreiben oder direkte Kontaktaufnahme durch Besuche. Dies wirkt sich selbstverständlich unmittelbar auf die Kosten aus, ist aber ethisch leichter vertretbar als differenzierte Incentives. Die Anzahl der Kontakte und die Art ist hier wesentlich. Durch erneute Anschreiben verringert sich die Chance, dass die Schreiben bspw. als Werbung eingestuft werden und nicht als Befragung wahrgenommen werden.
Insgesamt lässt sich nach zwei Pilotstudien festhalten, dass die Ausschöpfung im Vergleich zur Literatur nicht unterdurchschnittlich ist, aber durchaus Verbesserungspotentiale erkennbar sind. Die Rekrutierung über das ZMR ist als Ausgangspunkt vom Standpunkt der Datenqualität sicherlich geeigneter als die Rekrutierung aus dem Mikrozensus. Auch die bedingungslosen Incentives als Erstrekrutierung sind den konditionalen Incentives vorzuziehen. Ein nächster Schritt wäre der Versuch die Strategie der Kombination von Web- und postalischer Befragung von Messer und Dillman (2011) in Österreich umzusetzen und zu sehen, ob die Zusammensetzung der Stichprobe homogener wird.
Außerdem gilt es den Erfolg der Gruppe C aufzugreifen und zu validieren. Wie Edwards, Dillman und Smyth (2014) zeigen, hat die Verknüpfung von Umfragen mit Institutionen einen Effekt auf Befragungen, der für die Steigerung der Qualität und Ausschöpfung genutzt werden kann. Mit dem derzeitigen Design ist schwer abschätzbar, ob die Resultate für Gruppe C tatsächlich auf dem Folder und der Verknüpfung mit der Universität Wien und PUMA beruhen. Hier würde ein dezidiertes Design mit dem Vergleich über mehrere Gruppen Aufschluss geben.
Da die Literatur zeigt, dass für die Rekrutierung von Panelstudien Incentives unterschiedliche Wirkung zeigen, ist es empfehlenswert das Design der Gruppe D, welches die höchste Wiederholungsbereitschaft erzeugte, weiterzuverfolgen. Dies erscheint auch deshalb sinnvoll, da Gruppe C und D die besten beiden Optionen sind und in beiden Varianten im Aviso-Brief der Folder verwendet wurde.

8 Literaturverzeichnis

AAPOR. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th editio Aufl.). Autor.

Benesch, T. (2012). *Schlüsselkonzepte zur statistik: die wichtigsten methoden, verteilungen, tests anschaulich erklärt*. Springer-Verlag.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S. & Krieger, U. (2016, Februar). A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe. *Social Science Computer Review*, *34* (1), 8–25. doi: 10.1177/0894439315574825

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A. & Lavrakas, P. J. (2014). *Online panel research: A data quality perspective*. John Wiley & Sons, Ltd.

Callegaro, M., Manfreda, K. L. & Vehovar, V. (2015). *Web survey methodology*. Sage Publications.

de Heij, V., Schouten, B. & Shlomo, N. (2010, Mai). Risq manual tools in sas and r for the computation of r-indicators and partial r-indicators work package 8 deliverable 12.1 [Software-Handbuch].

Drewes, F. (2014). An empirical test of the impact of smartphones on panel-based online data collection. In *Online panel research* (S. 367–386). John Wiley & Sons, Ltd.

Edwards, M. L., Dillman, D. A. & Smyth, J. D. (2014). An experimental test of the effects of survey sponsorship on internet and mail survey response. *Public Opinion Quarterly*, *78* (3), 734–750.

Ernst Stähli, M. & Joye, D. (2016). Incentives as a Possible Measure to Increase Response Rates. *The SAGE Handbook of Survey Methodology*.

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, *1* (1), 58–70.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

Gumprecht, D. & Oismüller, A. (2013). Non-Response im Mikrozensus. *Statistische Nachrichten*, *11* (2013), 1046–1061.

Laurie, H. & Lynn, P. (2009). The use of respondent incentives on longitudinal surveys. *Methodology of longitudinal surveys*, 205–233.

Lumley, T. (2016). *survey: analysis of complex survey samples.* (R package version 3.31-2)

Lyberg, L. & Weisberg, H. (2016). *Total Survey Error: A Paradigm for Survey Methodology*. Sage Publications.

Messer, B. L. & Dillman, D. A. (2011). Surveying the general public over the internet using address-based sampling and mail contact procedures. *Public Opinion Quarterly*, *75* (3), 429–457.

Schoeni, R. F., Stafford, F., Mcgonagle, K. A. & Andreski, P. (2013). Response rates in national panel surveys. *The ANNALS of the American Academy of Political and Social Science*, *645* (1), 60-87.

Seymer, A. (2017). *Evaluierung und Dokumentation der Rekrutierungsexperimente bei der PUMA-Erhebung Q2/2016*. PUMA/Statistik Austria.

Singer, E. & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645* (1), 112–141.

Till, M. (2017). *PUMA Onlinebefragung Q4/2016 - Feldbericht*. Statistik Austria.

Ämter des Bundes und der Länder, S. (2011). *Zensus 2011: Vielfältiges Deutschland*. Information und Technik Nordrhein-Westfalen.

9 Anhang

9.1 Kreuztabellen zum Vergleich Brutto- und Nettostichprobe

Tabelle 19: Vergleich Brutto- und Nettostichprobe nach Incentivegruppen (ungewichtet)

	Nettostichprobe		
Incentivegruppe	Nein	Ja	Gesamt
Gruppe A kein incentive	420	59	479
	14,4%	8,2%	13,2%
Gruppe B 2 Euro im Säckchen	342	94	436
	11,7%	13,0%	12,0%
Gruppe C 2 Euro in Folder	374	114	488
	12,8%	15,8%	13,4%
Gruppe D 2 Euro in Folder + 5 Euro im Säckchen konditional	337	144	481
	11,6%	20,0%	13,2%
Gruppe E 5 Euro im Säckchen	300	85	385
	10,3%	11,8%	10,6%
Gruppe F 5 Euro Münze im Säckchen konditional bei Beantwortung	331	59	390
	11,4%	8,2%	10,7%
Gruppe G 10 Eurogutschein konditional bei Beantwortung	807	166	973
	27,7%	23,0%	26,8%
Gesamt	2911	721	3632
	100,0%	100,0%	100,0%

$\chi^2 = 63{,}851; df = 6; p < 0{,}001$

,

Tabelle 20: Vergleich Brutto- und Nettostichprobe nach Geschlecht (ungewichtet)

	Nettostichprobe		
Geschlecht	Nein	Ja	Gesamt
1	1400	356	1756
	48,1%	49,4%	48,3%
2	1511	365	1876
	51,9%	50,6%	51,7%
Gesamt	2911	721	3632
	100,0%	100,0%	100,0%

$\chi^2 = 0{,}381; df = 1; p = 0{,}537$

Tabelle 21: Vergleich Brutto- und Nettostichprobe nach Altersgruppen (ungewichtet)

	Nettostichprobe		
Altersgruppen	Nein	Ja	Gesamt
16 bis 24 Jahre	436	131	567
	15,0%	18,2%	15,7%
25 bis 34 Jahre	511	125	636
	17,6%	17,4%	17,6%
35 bis 44 Jahre	485	118	603
	16,7%	16,4%	16,7%
45 bis 54 Jahre	581	155	736
	20,0%	21,5%	20,3%
55 bis 64 Jahre	486	123	609
	16,8%	17,1%	16,8%
65 bis 74 Jahre	401	68	469
	13,8%	9,4%	13,0%
Gesamt	2900	720	3620
	100,0%	100,0%	100,0%

$\chi^2 = 12{,}96; df = 5; p = 0{,}024$

Tabelle 22: Vergleich Brutto- und Nettostichprobe nach Bildung (ungewichtet)

	Nettostichprobe		
Bildung	Nein	Ja	Gesamt
unbekannt oder maximal Pflichtschule	1079	160	1239
	37,1%	22,2%	34,1%
Lehre, berufsbildende Schule oder Matura	1559	410	1969
	53,6%	56,9%	54,2%
Hochschule	273	151	424
	9,4%	20,9%	11,7%
Gesamt	2911	721	3632
	100,0%	100,0%	100,0%

$\chi^2 = 104{,}855; df = 2; p < 0{,}001$

Tabelle 23: Vergleich Brutto- und Nettostichprobe nach Urbanisierung (ungewichtet)

	Nettostichprobe		
Urbanisierung	Nein	Ja	Gesamt
niedrig	905	224	1129
	31,1%	31,1%	31,1%
mittel	885	211	1096
	30,4%	29,3%	30,2%
hoch	1121	286	1407
	38,5%	39,7%	38,7%
Gesamt	2911	721	3632
	100,0%	100,0%	100,0%

$\chi^2 = 0{,}448$; $df = 2$; $p = 0{,}799$

Tabelle 24: Vergleich von Fragebogenabbruch und vollständigen Fragebögen nach Bildung (ungewichtet)

	Nettostichprobe		
Bildung	Nein	Ja	Gesamt
unbekannt oder maximal Pflichtschule	18	160	178
	36,0%	22,2%	23,1%
Lehre, berufsbildende Schule oder Matura	28	410	438
	56,0%	56,9%	56,8%
Hochschule	4	151	155
	8,0%	20,9%	20,1%
Gesamt	50	721	771
	100,0%	100,0%	100,0%

$\chi^2 = 7{,}764$; $df = 2$; $p = 0{,}021$

Tabelle 25: Kreuztabelle zum verwendeten Gerät und Bildung für die vollständigen Fragebögen (ungewichtet)

Gerät	Bildung			
	unbekannt oder maximal Pflichtschule	Lehre, berufsbildende Schule oder Matura	Hochschule	Gesamt
unbekannt	42	109	34	185
	26,2%	26,6%	22,5%	25,7%
PC	98	263	103	464
	61,3%	64,1%	68,2%	64,4%
Mobiltelefon	15	13	5	33
	9,4%	3,2%	3,3%	4,6%
Tablet	5	25	9	39
	3,1%	6,1%	6,0%	5,4%
Gesamt	160	410	151	721
	100,0%	100,0%	100,0%	100,0%

$\chi^2 = 13{,}666; df = 6; p = 0{,}034$

Tabelle 26: Kreuztabelle zum verwendeten Gerät und Bildung für die abgebrochenen Fragebögen (ungewichtet)

Gerät	Bildung			
	unbekannt oder maximal Pflichtschule	Lehre, berufsbildende Schule oder Matura	Hochschule	Gesamt
unbekannt	4	1	2	7
	22,2%	3,6%	50,0%	14,0%
PC	8	14	2	24
	44,4%	50,0%	50,0%	48,0%
Mobiltelefon	6	7	0	13
	33,3%	25,0%	0,0%	26,0%
Tablet	0	6	0	6
	0,0%	21,4%	0,0%	12,0%
Gesamt	18	28	4	50
	100,0%	100,0%	100,0%	100,0%

$\chi^2 = 12{,}959; df = 6; p = 0{,}044$