

Michael Ochsner

Systematische Methodenevaluation im Rahmen der Bevölkerungsbefragung der Stadt Zürich

Analyse und Empfehlungen

Im Auftrag der Stadtentwicklung Zürich

Lausanne, 15.11.2015

Inhaltsverzeichnis

Zusammenfassung	3
1 Einleitung	4
2 Ausführungen zum Design der Methodenevaluation	6
2.1 Beschreibung des Designs	7
2.2 Erläuterungen zum Design	9
3 Auswirkungen auf die Repräsentativität	10
3.1 Rücklaufquoten	10
3.2 Repräsentativität in Bezug auf die Stichprobe	12
3.3 Repräsentativität in Bezug auf die Population	15
4 Auswirkungen eines Mode-Wechsels auf die Verteilungen der Variablen	20
4.1 Allgemeine Erkenntnisse zur Vergleichbarkeit der CATI-Studie mit der Web/Papier-Kombination	21
4.2 Auswirkungen der Gewichtung	27
4.3 Konkrete Erläuterungen an Beispielen aus dem Bericht zur Bevölkerungsbefragung 2013	27
5 Empfehlungen	31
5.1 Umstellung des Erhebungsmodus auf Web/Papier	31
5.2 Zufallsstichprobe als Basis der Erhebung	31
5.3 Verzicht auf Gewichtung der Web/Papier-Kombination	32
5.4 Weiterführung der Zeitreihe mit zwei Werten für 2015	33
5.5 Verwendung von 95%-Vertrauensintervallen in Graphiken	33
5.6 Gründliche Evaluation der Zusammenarbeit mit einem Erhebungsinstitut resp. eigenständige Durchführung	34
5.7 Zeitpunkt des Wechsels des Erhebungsmodus 2019	34
5.8 Pragmatisches, aber breites Sprachenangebot	35
Appendix	37
Literatur	37
Danksagung	38

Zusammenfassung

Seit 1999 führt Stadtentwicklung Zürich alle zwei Jahre eine repräsentative telefonische Bevölkerungsbefragung über die Lebensbedingungen in der Stadt Zürich durch. In diesem Zeitraum ist die Ausschöpfung – d.h. der Anteil realisierter Interviews an der Ausgangsstichprobe – kontinuierlich von 54% im Jahre 1999 auf 36% im Jahre 2013 gesunken. Stadtentwicklung Zürich hat deshalb beschlossen, die Vor- und Nachteile eines Wechsels des Erhebungsmodus in einer systematischen Methodenevaluation mittels einer Parallelstudie experimentell abzuklären. Dieser Bericht fasst erste Erkenntnisse aus dieser Methodenevaluation zusammen und leitet daraus Empfehlungen zur Weiterführung der Bevölkerungsbefragung ab, mit besonderer Berücksichtigung der Aufrechterhaltung der Zeitreihe.

Die Methodenevaluation zeigt, dass der Wechsel von CATI zu einem anderen Erhebungsmodus nicht nur möglich, sondern sogar vorteilhaft ist. Insbesondere die Kombination eines Web- und Papierfragebogens weist gute Eigenschaften auf. So erreicht die Web/Papierkombination eine mindestens so gute Repräsentativität bezüglich der Bevölkerung wie die quotierte und gewichtete CATI-Studie, auch wenn eine reine Zufallsstichprobe und keine Gewichtung verwendet wird. Dies verringert den Aufwand und erhöht die Datenqualität für statistische Analysen.

In Bezug auf die Auswirkungen eines Wechsels auf die Resultate der Studie und die Zeitreihe zeigen die Analysen, dass zwar bei vielen Variablen ein Effekt zu verzeichnen ist, dieser Effekt aber in allen Fällen klein ausfällt und eine Interpretation der Resultate nicht gänzlich verändern würde. Die erreichte Stichprobengrösse bei der Web/Papier-Kombination ermöglicht es ausserdem, relativ zuverlässige Schätzer für die Bevölkerungsbefragung 2015 zu liefern, so dass eine Weiterführung der Zeitreihe nahtlos möglich ist, indem man für die Jahre 1999 bis 2015 die Werte der CATI-Studien vergleicht und ab dem Jahr 2015 die Werte der Web/Papier-Kombination. Innerhalb 2015 wird ersichtlich, wie sich der Effekt des Wechsels von CATI auf Web/Papier auswirkt.

In der Regel fallen die Effekte dahingehend aus, dass die Werte etwas kritischer sind: Dadurch, dass der Anteil der Personen, die an der Web/Papier-Kombination teilgenommen haben, deutlich höher ist als jener, die an der CATI-Studie teilgenommen haben (56% vs. 33% Rücklaufquote), sind mehr unzufriedene Personen in der Stichprobe vertreten, was den Schätzer etwas realistischer macht (die Gewichtung der CATI-Studie wirkt ebenfalls in diese Richtung).

Es lässt sich schlussfolgern, dass ein Wechsel zu einer Web/Papier-Kombination verschiedene Vorteile aufweist und dass der Zeitpunkt des Wechsels aufgrund eines Wechsels vom Zweijahres- zum Vierjahresrhythmus günstig ist. Empfehlungen zur Umsetzung eines solchen Wechsels schliessen den Bericht ab.

1 Einleitung

Seit 1999 führt Stadtentwicklung Zürich alle zwei Jahre eine repräsentative telefonische Bevölkerungsbefragung über die Lebensbedingungen in der Stadt Zürich durch. In diesem Zeitraum ist die Ausschöpfung – d.h. der Anteil realisierter Interviews an der Ausgangsstichprobe – kontinuierlich von 54% im Jahre 1999 auf 36% im Jahre 2013 gesunken, also ziemlich genau um ein Drittel. Da die Ausschöpfung einer der wichtigsten Indikatoren für die Güte einer Befragung ist, gefährdet ein solcher Rückgang die Aussagekraft der ganzen Erhebung.

Die allgemein kolportierte Annahme, hinter gesunkenen Ausschöpfungsquoten stecke eine allgemeine Befragungsmüdigkeit oder -skepsis, lässt sich im Falle der Zürcher Befragung nicht bestätigen: Der Anteil der Verweigerungen hat sich seit 1999 nämlich nahezu halbiert (von 11% auf 6%).

Was sich hingegen erheblich verschlechtert hat, ist die Erreichbarkeit per Telefon: Waren es 1999 rund 17% der Ausgangsadressen, für die keine brauchbare Telefonnummer eruiert werden konnte, so stieg dieser Anteil in der jüngsten Befragung 2013 auf 34%. Das bedeutet eine Verdoppelung in nur 14 Jahren (alle Zahlenangaben: Bevölkerungsbefragung 2013: Stadtentwicklung Zürich, 2013, S. 55).

Solche Probleme sind der Grund dafür, dass wissenschaftliche Befragungen schon seit einiger Zeit kaum mehr rein telefonisch durchgeführt werden. Natürlich muss der Übergang zu anderen oder gemischten Befragungsmodi (mixed mode interviewing) gut überlegt und zuverlässig abgeklärt werden. Stadtentwicklung Zürich hat deshalb beschlossen, die Vor- und Nachteile eines solchen Übergangs mittels einer systematischen Methodenevaluation experimentell abzuklären und für die optimale Realisierung der Methodenevaluation externe Beratung beizuziehen. FORS wurde damit beauftragt, in Zusammenarbeit mit der Stadtentwicklung Zürich ein Design für eine systematische Methodenevaluation zu entwickeln, das mögliche Effekte eines Wechsels des Erhebungsmodus abzuschätzen ermöglicht, und im Anschluss an die Erhebung die resultierenden Daten zu analysieren und Empfehlungen bezüglich möglicher Erhebungsmodi sowie allfälliger Kombinationen von Modi für die kommenden Erhebungen auszuarbeiten. Dabei soll ein besonderes Augenmerk auf die Auswirkungen eines solchen Wechsels auf die Zeitreihe der Daten der Bevölkerungsumfrage gelegt werden, die seit 1999 in regelmässigen Abständen durchgeführt wird.

Der Report ist folgendermassen gegliedert: Zuerst wird das gewählte Studiendesign beschrieben. Anschliessend folgen Informationen über die Durchführung der Studie. Es folgen dann Analysen zu Auswirkungen der verschiedenen Erhebungsmodi auf die Repräsentativität. Der nächste Abschnitt widmet sich den

Effekten, die ein Wechsel des Erhebungsmodus auf die Resultate der Bevölkerungsstudie haben könnte, mit einem Fokus auf die Kombination von Web- und Papiermodus. Der Bericht schliesst mit Empfehlungen zu einem allfälligen Wechsel des Befragungsmodus ab. Dabei wird im Report so weit wie möglich auf statistische Details verzichtet und eine möglichst verständliche Darstellung der bisweilen komplexen Zusammenhänge angestrebt.

2 Ausführungen zum Design der Methodenevaluation

Die letzten Ausgaben der Bevölkerungsbefragung der Stadt Zürich haben gezeigt, dass die Sicherstellung der Repräsentativität der telefonischen Befragungen (nachfolgend mit CATI bezeichnet) je länger je mehr mit grossem Aufwand verbunden ist, da der Anteil der Personen, die über eine öffentlich gelistete Telefonnummer erreicht werden können, in den letzten Jahren stetig abnahm und noch weiter abnehmen wird. Zusätzlich sind die Personen immer schlechter telefonisch erreichbar, was den Aufwand, Respondenten zu kontaktieren, nochmals erhöht. Dies zeigt sich in der steigenden Anzahl an Kontaktversuchen, die für ein Interview nötig sind. Deshalb stellt sich die Frage, ob ein Wechsel des Befragungsmodus Abhilfe schaffen könnte. Da für die Bevölkerungsbefragung der Stadt Zürich eine relativ lange Zeitreihe besteht (seit 1999), ist es wichtig, dass Effekte des Moduswechsels auf die Zeitreihe wissenschaftlich getestet werden. Dies setzt ein sorgfältiges, experimentelles Design der Methodenevaluation voraus. Folgende Fragen stehen dabei im Zentrum: Welchen Einfluss hat ein Modewechsel auf die Repräsentativität der Studie? Welche Effekte hat ein Modewechsel auf die Zeitreihe; kann die Vergleichbarkeit aufrecht erhalten werden? Ist eine Kombination von Modes sinnvoll?

Natürlich sind mit der Methodenevaluation grundsätzlich hohe Kosten verbunden, somit muss beim Design darauf geachtet werden, dass möglichst viele Alternativen bei möglichst geringem finanziellen Aufwand getestet werden können. Deshalb wurden der Stadtentwicklung drei Designs vorgeschlagen, die unterschiedliche Schwerpunkte haben. Insbesondere gilt es, dem Trade-Off zwischen dem Testen verschiedener Modekombinationen und der Verlässlichkeit der Schätzer bezüglich der Zeitreihe Beachtung zu schenken: Je mehr Modes und Kombinationen getestet werden, desto kleiner fällt die jeweilige Stichprobengrösse aus und desto unpräziser werden die Schätzer der Effekte auf die Vergleichbarkeit der Zeitreihe.

Gemeinsam mit der Auftraggeberin wurde ein Design ausgewählt, das die zwei vielversprechendsten Modekombinationen beinhaltet und aufgrund der Anlage der Kombinationen zumindest teilweise eine Abschätzung der Selektionseffekte und der Methodeneffekte der verschiedenen Erhebungsmodi erlaubt. Zudem erlaubt es auch die unveränderte Weiterführung der CATI-Erhebung, damit die Aufrechterhaltung der Zeitreihe gewährleistet ist. Abbildung 1 zeigt die schematische Anlage des Designs. Im Folgenden wird zuerst kurz das Design beschrieben und dann bezüglich der sich eröffnenden Analysemöglichkeiten erläutert.

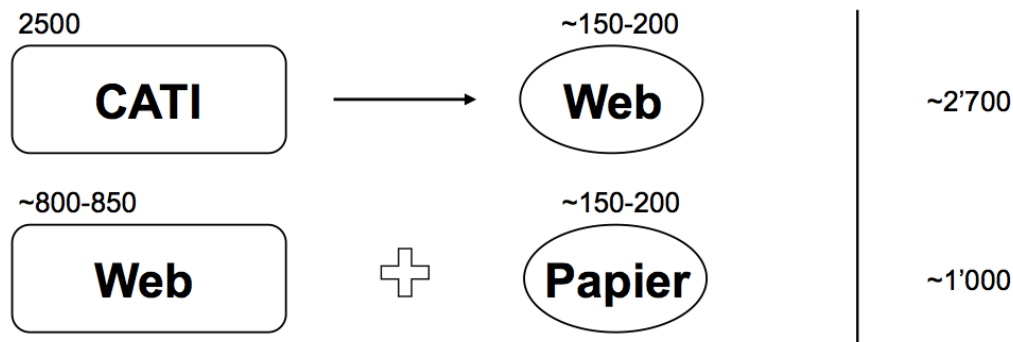


Abbildung 1. Schematische Darstellung des Designs der Methodenevaluation für die Bevölkerungsbefragung der Stadt Zürich

2.1 Beschreibung des Designs

Im Zentrum steht auch bei dieser Durchführung der Bevölkerungsumfrage die CATI-Befragung. Sie wird genau gleich weitergeführt wie in den vorigen Jahren. Nach Ende der Feldphase, also dem Erreichen von 2500 Respondenten per Telefon, wird jedoch zusätzlich den restlichen Stichprobenmitgliedern eine Einladung zu einem Websurvey zugeschickt (nachfolgend als Web-Follow-Up bezeichnet).

Unabhängig von der CATI/Web-Follow-Up Kombination wird basierend auf einer eigenen Stichprobe eine Parallelstudie durchgeführt, die einem Web- und Papier-Mixed-Mode Design folgt. Dabei wird den Personen in der Stichprobe ein Brief mit den Zugangsdaten zu dem Websurvey zugestellt. Der Brief enthält gleichzeitig die Information, dass auch ein Papierfragebogen bestellt werden kann. Letzterer wird dann für jene Personen, für die keine Telefonnummer auffindig gemacht werden konnte, zusätzlich dem zweiten Erinnerungsschreiben beigelegt. Die Personen, für welche eine Telefonnummer gefunden werden konnte, werden telefonisch kontaktiert. Ihnen wird je nach bevorzugtem Teilnahmemodus ein Email mit den Login-Daten oder ein Brief mit dem Papierfragebogen zugestellt. Es handelt sich also nicht um ein sequentielles Mixed-Mode-Design. Den Befragten wird von Beginn weg die Wahl des Modus gelassen, wobei allerdings der Aufwand, einen Papierfragebogen zu bestellen etwas grösser ist, als die Zugangsdaten im Webbrowser einzugeben.

Die Stichprobengrössen wurden vom Erhebungsinstitut so festgelegt, dass die Anzahl Respondenten im veranschlagten Rahmen zu liegen kommen (siehe Abbildung 1): So sollen aus der CATI/Web-Follow-Up Kombination rund 2'700 Interviews erfolgen, während aus der Parallelstudie zirka 1'000 Interviews resultieren sollten, bei einem erfahrungsgemässen Anteil von 25-30% Papierfragebögen.

Die Stichprobe wurde für die CATI/Web-Follow-Up Kombination genau so gezogen, wie sie in den letzten Jahren für die Bevölkerungsumfrage gezogen wurde: Stadt Zürich – Abteilung Organisation und Informatik / Applikationsentwicklung (OIZ) – zieht auf der Grundlage der Einwohner/Innendatenbank und gemäss der aktuellen Bevölkerungsstruktur eine geschichtete Zufallsstichprobe der in der Stadt Zürich wohnhaften und gemeldeten volljährigen Personen mit Schweizer Bürgerrecht oder einer B- oder C-Bewilligung, inkl. Wochenaufenthalter. Dieses Jahr umfasste die Stichprobe 9'103 Adressen, die in eine Basis (n=5'748) und eine Reservestichprobe (n=3'355) aufgeteilt wurden. Die Personen in der Basisstichprobe wurden bei Studienbeginn angeschrieben und anschliessend kontaktiert. Von der Reservestichprobe wurden im Anschluss nur jene Quotenzellen aktiviert, für die noch nicht genügend Interviews geführt werden konnten.

Die Stichprobe der Parallelstudie (Web/Papier) wurde ebenfalls von der OIZ basierend auf der selben Grundgesamtheit gezogen. Allerdings wurde hier eine reine Zufallsstichprobe ohne Quotierung gezogen. Dies hat verschiedene Gründe. Einerseits führt eine Quotierung potenziell zu Verzerrung in der Repräsentativität bezüglich anderer Variablen, die nicht in die Quotierung einfließen, weshalb FORS der Stadtentwicklung Zürich empfohlen hat, bei einem Wechsel von CATI auf Web/Papier eine reine Zufallsstichprobe zu verwenden. Andererseits eignet sich eine solche experimentelle Parallelstudie gerade auch dafür zu testen, ob mit einer unquotierten Stichprobe eine genügende Anzahl an Antworten pro Quotenzelle erreicht werden kann, sodass Analysen auch für Subpopulationen möglich bleiben.

Die CATI Befragung wird wie in den Vorjahren in verschiedenen Sprachen angeboten (Deutsch, Italienisch, Englisch, Portugiesisch, Spanisch und Serbokroatisch). Aus Kostengründen wurde bei der Parallelstudie darauf verzichtet, mehrere Sprachen zu verwenden. Die Web- und Papierfragebögen werden nur in Deutsch angeboten. Dies hat leider den Effekt, dass die Vergleichbarkeit zwischen CATI und Web/Papier leidet: Zu Methoden- und Selektionseffekten kommt zusätzlich der Effekt verschiedener Sprachangebote. Allerdings wurden in der Vergangenheit nur sehr wenige Interviews in einer anderen Sprache als Deutsch geführt (9.9%, davon allein 7% Italienisch und Englisch, s. Bevölkerungsbefragung 2013: Stadtentwicklung Zürich, 2013), somit dürfte der Spracheneffekt nicht allzu stark ausfallen. Aus dem Angebot nur einer Sprache erwächst andererseits ein weiterer Nutzen der Parallelstudie: Da das Angebot vieler Sprachen mit hohem Aufwand und nicht vernachlässigbaren Kosten verbunden ist, kann geprüft werden, ob auch nur mit der Befragungssprache Deutsch eine ansprechende Repräsentativität unter Ausländern erreicht werden kann.

2.2 Erläuterungen zum Design

Die Durchführung der unveränderten CATI-Studie gewährleistet, dass auch während des Austestens anderer Erhebungsmodi die Zeitreihe unverändert fortgeführt wird. Zweitens ermöglicht sie den Vergleich der alternativen Erhebungsmodi mit der bisherigen Methode im gleichen Jahr.

Die Parallelstudie beinhaltet eine vielversprechende Alternative zur CATI-Studie: Die Kombination Web/Papier wird häufig verwendet und zeichnet sich einerseits dadurch aus, dass die Resultate beider Modi gut vergleichbar sind und Unterschiede bisher selten festgestellt wurden; andererseits zeichnet sie sich dadurch aus, dass beide Modi unterschiedliche Personen ansprechen: Den Webfragebogen füllen generell eher junge Personen und Männer aus, während ältere Personen und Frauen tendenziell den Papierfragebogen bevorzugen. Die Kombination dieser beiden Eigenschaften führt dazu, dass Web/Papier-Befragungen grundsätzlich eine gute Repräsentativität erreichen. Die Stichprobengröße ist bei der Parallelstudie so angelegt, dass die Web/Papier-Kombination genügend Respondenten erzeugt, damit sie auch verlässliche Aussagen über Effekte auf die Zeitreihe der CATI-Befragung zulässt.

Das Design informiert auch über das Potential einer CATI-Studie mit nachgeschaltetem Webfragebogen, jene Personen zu erreichen, die nicht per Telefon antworten wollen. Aus Kostengründen wurde das Design etwas vereinfacht, indem auf eine separate CATI/Web-Parallelstudie verzichtet wurde. Stattdessen wurde der unveränderten CATI-Studie ein Web-Follow-Up nachgeschaltet. Aus dieser Kombination kann dennoch erschlossen werden, wie sich eine solche Kombination auf die Repräsentativität und auch auf die Resultate auswirkt.

3 Auswirkungen auf die Repräsentativität

Ein erstes Mass, das häufig mit der Güte einer Befragung in Verbindung gebracht wird, ist die Rücklaufquote. Dieses Mass ist aber in Bezug auf die Beurteilung der Repräsentativität nicht nur bedingt tauglich, sondern gibt eher Auskunft über das potentielle Risiko von Verzerrung sowie über die Anstrengungen, die nötig sind, um Respondenten für eine Teilnahme zu gewinnen. Folglich wird nur kurz auf die Rücklaufquoten eingegangen. Aussagekräftiger ist die Untersuchung der Repräsentativität bezüglich einiger Variablen in der Stichprobe und der Population, die deshalb im Folgenden ausführlicher behandelt wird.

Da die CATI/Web-Follow-Up-Studie und die Web/Papier-Parallelstudie auf zwei separaten Stichproben basieren, können die Auswirkungen verschiedener Modes auf die Repräsentativität auf zwei verschiedene Arten untersucht werden: einerseits in Bezug auf die Stichprobe, andererseits in Bezug auf die Population. Die Analyse in Bezug auf die Stichprobe gibt darüber Auskunft, welche Verzerrungen mit den untersuchten Modes verbunden sind. Eine Analyse in Bezug auf die Population könnte Verzerrungen verdecken, denn vermeintliche Unterschiede könnten auf unterschiedliche Zusammensetzungen der Stichproben zurückzuführen sein oder unterschiedliche Zusammensetzungen der Stichproben könnten dazu führen, dass eigentliche Verzerrungen nicht als solche erkannt würden. Die Analyse in Bezug auf die Population bietet dafür einen anderen Vorteil. Wenn sich, wie im vorliegenden Fall, die Art der Stichprobenziehung unterscheidet (die CATI/Web-Follow-Up-Studie basiert auf einer geschichteten Zufallsstichprobe während die Web/Papier-Paralleluntersuchung auf einer reinen Zufallsstichprobe basiert), zeigt eine Analyse in Bezug auf die Population die gesamte Repräsentativität auf, nämlich die der Kombination des Modes mit der Stichprobenart. Deshalb wird im Folgenden die Repräsentativität im Hinblick auf beide Bezugsgrößen untersucht.

3.1 Rücklaufquoten

Die Rücklaufquote, definiert als der Anteil der Personen aus der Stichprobe, die an der Befragung teilnehmen, wird häufig mit der Güte einer Befragung in Verbindung gebracht. Würden nämlich alle Personen, die Teil der Stichprobe sind, den Fragebogen beantworten, wären Repräsentativität und insbesondere auch optimale Voraussetzungen für statistische Analysen gegeben. In der Realität nehmen aber nie alle Personen der Stichprobe an der Untersuchung teil. Meist wird deshalb die Rücklaufquote als Qualitätskriterium der Studie und implizit auch als Indikator für die Repräsentativität verwendet. Hier gilt es zu bemerken, dass die Höhe der Rücklaufquote (abgesehen von einer Rücklaufquote nahe 100%) nicht zwingend mit Repräsentativität, geschweige denn mit der Qualität

der Daten, in Zusammenhang steht. Groves (2006) zeigt eindrücklich, dass auch Befragungen mit tiefer Rücklaufquote (15-20%) gute Repräsentativität leisten können, während Befragungen mit hohen Rücklaufquoten (70% und mehr) beträchtliche Verzerrungen aufweisen können. Nichtsdestotrotz nimmt das Risiko von Verzerrungen mit sinkenden Rücklaufquoten zu, insbesondere wenn gewisse Bevölkerungsgruppen systematisch nicht mehr teilnehmen (z.B. wenn sie keinen Festnetzanschluss mehr besitzen).

Deshalb werden im Folgenden als erstes Indiz für potentielle Verzerrungen, mindestens aber für den steigenden Aufwand, Respondenten zur Teilnahme zu bewegen, die Rücklaufquoten der verschiedenen Teilstudien präsentiert (siehe Tabelle 1).

Bei der CATI-Studie setzt sich der Trend sinkender Rücklaufquoten auch 2015 fort. Da als Minimum 2'500 Teilnehmer festgelegt wurden und es sich um eine quotierte Stichprobe handelt und alle Adressen für die Kontaktaufnahme verbraucht wurden, liegt die Rücklaufquote beim festgelegten Minimum von 33%. Jene Personen aus der Reservestichprobe, die aus einer Quotenzelle stammen, für welche die erforderliche Anzahl Interviews bereits erreicht wurde, wurden gar nicht erst kontaktiert. Wird die Gesamtstichprobe berücksichtigt, also auch jene Personen einbezogen, die aus Quotierungsgründen gar nicht kontaktiert wurden, obwohl sie in der Stichprobe gewesen wären, resultiert gar eine Rücklaufquote von 28%. Der Web-Follow-Up erreichte eine Rücklaufquote von 11%. Hier gilt es zu berücksichtigen, dass die Einladung zum Web-Follow-Up nur jenen Personen zugestellt wurde, die zwar per Brief zur CATI-Studie eingeladen wurden, von denen aber keine Telefonnummer bekannt war und die somit telefonisch nicht kontaktiert werden konnten. Für eine Befragung von Umfrageverweigerern (allerdings still verweigernd, da die Zielpersonen der Web-Follow-Up-Studie zwar nicht an der CATI-Befragung teilgenommen haben, sich jedoch auch nicht explizit verweigert haben) liegt die Rücklaufquote im Rahmen der Erwartungen. Für die Kombination CATI/Web-Follow-Up ergibt sich eine gesamte Rücklaufquote von 37%.

Den Onlinefragebogen füllten 42% der Personen in der Stichprobe der Parallelstudie aus. Weitere 14% wählten den Papierfragebogen. Dies resultiert in einer gesamten Rücklaufquote von 56% für die Kombination Web/Papier. Dies liegt über den Erwartungen, weshalb für die Stichprobe der Parallelstudie die Reservestichprobe nicht aktiviert werden musste. Zudem liegt das Verhältnis zwischen Web und Papier in etwa im erwarteten Bereich: 75% der Teilnehmenden wählten den Webfragebogen, 25% den Papierfragebogen.

Die Rücklaufquoten liefern einen Hinweis darauf, dass die Zürcher Bevölkerung besser mittels Web- und Papierfragebogen erreicht werden kann als mittels te-

lefonischer Befragungen. Der Aufwand für die 2'500 Telefoninterviews war dieses Jahr sehr gross, während in der Parallelstudie mit Web- und Papierfragebogen die Reservestichprobe nicht aktiviert werden musste. Als nächstes interessiert nun, ob die Bevölkerung mit den Alternativmodi ebenso gut repräsentiert ist wie mit der bisherigen Methode.

Tabelle 1. Rücklaufquoten per Mode und Mode-Kombination

Mode	CATI	Web	CATI/WebFU	Web	Papier	Web/Papier
Samplegrösse	7591	2840	7591	2100	2100	2100
Teilnehmende	2501	311	2812	885	286	1171
Rücklaufquote	33%	11%	37%	42%	14%	56%

Bemerkung. CATI: Computer-Assisted Telephone Interview; WebFU: Web-Follow-Up; Web: Online Fragebogen; Papier: Papierfragebogen

3.2 Repräsentativität in Bezug auf die Stichprobe

Um die Repräsentativität in Bezug zur Stichprobe zu untersuchen, sind zusätzliche Informationen zu den Personen in der Stichprobe nötig. Die Adressdatei der Stichprobe wurde mit einigen solchen Variablen geliefert. Dies sind namentlich die drei Variablen, die für die Quotierung der CATI-Studie verwendet wurden: Quartier (23 Stadtteile), Geschlecht und AusländerIn/SchweizerIn resp. die Art der Aufenthaltsbewilligung (SchweizerIn, B- oder C-Bewilligung). Zusätzlich wurden die Haushaltsgrösse, das Alter und die Anzahl Jahre, die die Zielperson in der Stadt Zürich wohnt, mitgeliefert.

Um die Repräsentativität in Bezug zur Stichprobe im Hinblick auf diese sechs Variablen zu untersuchen, wurde für jeden Mode und die beiden Mode-Kombinationen je ein logistisches Modell berechnet, das die Teilnahme der Zielperson an der Untersuchung mit den sechs Variablen als erklärende Variablen modelliert. Zwei Masse dienen der Abschätzung der Repräsentativität: Als erstes Mass wird das Signifikanzniveau der erklärenden Variablen verwendet, das darauf hinweist, ob eine Variable die Teilnahme der Zielperson beeinflusst. Als zweites Mass wird die durchschnittliche prozentuale Abweichung der vorhergesagten Wahrscheinlichkeit (sog. predicted probabilities) von der Rücklaufquote verwendet. Dieses zweite Mass dient dazu abzuschätzen, wie gross die Verzerrung bezüglich der Stichprobe ist (vgl. ein ähnliches Mass für Repräsentativität in Lipps & Ochsner, 2016). Es wird folgendermassen berechnet: Nach der Schätzung des logistischen Modells wird für jede Beobachtung die durch die Variablen bedingte Wahrscheinlichkeit der Teilnahme berechnet. Anschliessend wird für jede Ausprägung der Variablen berechnet, wie viele Personen mit dieser Ausprägung an der Studie teilnehmen sollten (in anderen Worten, die Rücklaufquote, die aufgrund der Variablenausprägung erwartet wird). Anschliessend

wird diese „erwartete Rücklaufquote“ von der erreichten Rücklaufquote abgezogen. Für jede Variable wird dann die durchschnittliche prozentuale Abweichung von der Rücklaufquote berechnet. Dieses Mass zeigt, wie gross die Verzerrung bezüglich dieser Variable ist. Wird dieses Mass über alle Variablen summiert, ergibt sich ein Mass für die Verzerrung des Modes bezüglich dieser Variablen in der Stichprobe.

Tabelle 2 fasst die Resultate bezüglich des ersten Masses zusammen, das zeigt, welche Variablen einen signifikanten Einfluss auf die Teilnahmewahrscheinlichkeit der Zielpersonen haben. Es wird ersichtlich, dass durchaus Unterschiede zwischen den verschiedenen Modes bestehen. Nimmt man ein Signifikanzniveau von $p < 0.05$ als kritisches Mass, zeigt sich, dass sämtliche Variablen einen Einfluss auf die Teilnahmewahrscheinlichkeit in der CATI-Studie haben. Wird die CATI-Studie allerdings mit dem Web-Follow-Up kombiniert, erweist sich der Einfluss zweier Variablen, nämlich des Geschlechts und die Anzahl Jahre, die die Zielperson in der Stadt Zürich wohnt (kurz: Zuzug), als nicht mehr signifikant. Im Vergleich dazu zeigt sich, dass in der Kombination Web/Papier zusätzlich zu diesen beiden Variablen auch noch das Alter keinen signifikanten Einfluss mehr auf die Teilnahmewahrscheinlichkeit ausübt. Interessant dabei ist, dass das Alter sowohl beim Websurvey wie auch beim Papierfragebogen separat einen hochsignifikanten Einfluss auf die Teilnahmewahrscheinlichkeit hat – am Websurvey nehmen mehr junge Personen teil, beim Papierfragebogen sind ältere Personen überrepräsentiert. Der Effekt hebt sich aber in der Kombination auf.

Tabelle 2. Signifikanz der erklärenden Variablen des logistischen Modells für die Teilnahme an der Bevölkerungsbefragung für verschiedene Modes und Mode-Kombinationen

	CATI	WebFU	CATI/WebFU	Web	Papier	Web/Papier
Permit B,C,CH	0.000	0.351	0.000	0.004	0.767	0.003
Geschlecht	0.042	0.278	0.188	0.399	0.002	0.205
Stadtteil	0.000	0.188	0.000	0.093	0.013	0.04
Alter	0.000	0.090	0.000	0.000	0.000	0.208
Zuzug	0.028	0.090	0.154	0.263	0.524	0.104
Haushaltsgrösse	0.004	0.006	0.004	0.009	0.155	0.000

Der Nachteil dieser Methode ist, dass einerseits das Signifikanzniveau abhängig von der Stichprobengrösse ist, weshalb auch nicht auf die einzelnen Modes mit kleiner Stichprobengrösse (Web-Follow-Up, Web und Papier) eingegangen wurde, und dass andererseits das Ausmass der Verzerrung nicht ersichtlich

wird. Deshalb wird in Tabelle 3 das zweite Mass, die durchschnittliche prozentuale Abweichung der vorhergesagten Wahrscheinlichkeit (also der „erwarteten“ Rücklaufquote) von der erreichten Rücklaufquote präsentiert.

Tabelle 3. Durchschnittliche prozentuale Abweichung der vorhergesagten Teilnahme-wahrscheinlichkeit von der Rücklaufquote

	CATI	WebFU	CATI/WebFU	Web	Papier	Web/Papier
Permit B,C,CH	20.9	7.6	16.7	11.3	6.7	9.4
Geschlecht	3.2	5.9	2.0	2.1	16.1	2.5
Stadtteil	12.7	23.2	12.3	11.7	30.3	9.6
Alter	15.8	19.5	9.4	16.9	43.4	4.5
Zuzug	5.5	17.6	3.6	5.7	8.9	5.2
Haushaltsgrösse	5.7	25.5	6.7	6.4	22.0	8.9
Summe	63.6	99.2	50.6	54.2	127.4	40.1

Aus Tabelle 3 wird ersichtlich, dass die Kombination Web/Papier bezüglich fast aller Variablen deutlich weniger Abweichungen aufweist als die anderen Modes und Mode-Kombinationen. Einzig die Haushaltsgrösse ist in der Web/Papier-Kombination etwas grösser als bei der CATI-Studie. Werden die prozentualen Abweichungen über alle Variablen summiert, zeigt sich, dass die CATI-Studie durch den Web-Follow-Up bezüglich der Repräsentativität zur Stichprobe etwas optimiert wird. Am wenigsten Abweichungen zur Stichprobe weist die Web/Papier-Kombination auf. Auch hier wird wieder ersichtlich, dass bezüglich des Alters sowohl die Web-Studie als auch insbesondere die Papier-Studie einige Verzerrungen aufweisen, die Kombination aber zu einer ziemlich guten Altersverteilung führt.

Die Vorgehensweise, ein logistisches Modell als Basis zu verwenden, hat den Vorteil, dass die konditionale Verzerrung abgeschätzt werden kann. Das heisst, dass Personen, die zur Verzerrung beim Alter führen und gleichzeitig auch zur Verzerrung bei der Haushaltsgrösse beitragen, nicht bei beiden Variablen zu Buche schlagen. Die Verzerrungen in Tabelle 3 sind also immer „unter Kontrolle der anderen Variablen“ zu verstehen. Werden nur die prozentualen Abweichungen der Verteilung einer Variable unter den Teilnehmenden mit der prozentualen Verteilung derselben Variable in der Stichprobe verglichen, fällt der Unterschied entsprechend grösser aus. Tabelle 4 zeigt die unktionale Verzerrung pro Variable. Konkret wurde die erreichte Rücklaufquote für jede Kategorie einer Variable berechnet und von der Gesamtrücklaufquote subtrahiert. Diese Abweichung wurde durch die Gesamtrücklaufquote geteilt, um die prozentuale Abwei-

chung der erreichten Rücklaufquote einer Antwortkategorie von der erwarteten Rücklaufquote zu erhalten.

Tabelle 4. Durchschnittliche prozentuale Abweichung der erreichten von der erwarteten Rücklaufquote per Variablenkategorie (unkonditionale Verzerrung)

	CATI	WebFU	CATI/WebFU	Web	Papier	Web/Papier
Permit B,C,CH	29.0	39.3	21.9	8.7	16.2	10.5
Geschlecht	7.1	13.5	4.8	3.3	20.7	2.6
Stadtteil	15.6	20.3	14.4	11.5	34.2	10.5
Alter	22.0	55.5	13.5	14.8	48.1	4.3
Zuzug	18.8	53.6	10.8	8.4	34.9	5.7
Haushaltsgrösse	7.7	31.2	10.0	6.3	38.9	9.6
Summe	100.1	213.4	75.4	53.0	193.0	43.2

3.3 Repräsentativität in Bezug auf die Population

Stichproben können gegenüber der Population verzerrt sein. Deshalb interessiert insbesondere, ob ein Set von Teilnehmenden bezüglich der Population repräsentativ ist. Im Unterschied zur Stichprobe sind aber die Variablen nicht für alle Personen der Population vorhanden. Deshalb kann nur die unkonditionale Verzerrung der einzelnen Modes untersucht werden. Im Folgenden wird also die durchschnittliche prozentuale Abweichung der Verteilung einer Variable unter den Teilnehmenden von der Verteilung derselben Variable in der Population untersucht (entsprechend Tabelle 4 der Repräsentativität in Bezug auf die Stichprobe).

Leider wurden in den Fragebögen nicht viele Variablen erfasst, für welche die Verteilung in der Population bekannt ist. Deshalb beschränkt sich diese Untersuchung einerseits auf die Variablen, die in der CATI-Studie für die Quotierung verwendet wurden (AusländerIn/SchweizerIn, Geschlecht, Stadtteil), und andererseits auf Variablen, für welche die Verteilung in der Population ausfindig gemacht werden konnte (Alter, Bildung und Nationalität). Tabelle 5 zeigt die Verzerrung in Bezug auf die Population.

Auch in Bezug auf die Population erweist sich die Kombination Web/Papier als überlegen. Besonders überraschend ist, dass sogar in Bezug auf die Quotierungsvariablen die Web/Papier-Kombination leicht besser abschneidet als das CATI-Design. Es sei daran erinnert, dass bereits die Stichprobe geschichtet gezogen wurde und von der Reserve nur diejenigen Personen aktiviert wurden, die in einer unterrepräsentierten Quotenzelle lagen. Trotz dieser Massnahmen

weist die Web/Papier-Kombination weniger sozio-demographische Verzerrung auf, obwohl bei dieser Kombination keine Quotierung erfolgte¹.

Tabelle 5. Durchschnittliche prozentuale Abweichung der erreichten von der erwarteten teilnehmenden Personen per Variablenkategorie

	CATI	WebFU	CATI/WebFU	Web	Papier	Web/Papier
Ausländer	3.9	66.3	10.8	9.6	14.4	10.7
Geschlecht	3.2	23.7	5.4	4.5	18.8	1.2
Stadtteil	23.4	31.6	23.7	13.8	34.3	12.6
<i>Quota</i>	<i>29.6</i>	<i>91.3</i>	<i>32.7</i>	<i>28.5</i>	<i>57.4</i>	<i>25.2</i>
Alter	6.1	12.9	4.0	5.3	16.1	1.1
Bildung	4.2	16.4	5.6	9.8	9.5	5.2
Nationalität	17.0	60.0	22.2	21.5	25.2	20.0
Summe	87.3	302.2	104.3	93.0	175.8	76.1
Summe ohne Quota und Ausländer	53.8	144.6	60.9	55.0	104.0	40.2

Bemerkung. Quota bezieht sich auf die Interaktion der Variablen Ausländer, Geschlecht und Stadtteil. Die durchschnittliche prozentuale Abweichung wurde per Quotenzelle berechnet.

Die Web/Papier-Kombination zeigt allerdings eine relativ starke Verzerrung bezüglich Ausländern auf: Es sind zu viele Schweizer unter den Teilnehmenden. Es fällt aber auf, dass sobald man die Nationalität etwas besser aufschlüsselt (die häufigsten 17 Nationalitäten und die restlichen Nationalitäten nach Kontinenten aufgeteilt), die Repräsentativität der Web/Papier-Kombination nicht mehr deutlich schlechter als die der CATI-Studie ist (Ausländer: 10.7 vs. 3.9 und Nationalität: 20.0 vs. 17.0). Die zweite Variable, bei welcher die Web/Papier-Kombination schlechter als CATI abschneidet, ist die Bildung: Personen mit einer Tertiärausbildung sind unterrepräsentiert (7.8 Prozentpunkte), während Personen mit einer Sekundarbildung überrepräsentiert sind (6.7 Prozentpunkte). Personen mit einem Grundschulabschluss sind nur leicht überrepräsentiert (1.2 Prozentpunkte). Die durchschnittliche Abweichung bezüglich der Bildung ist allerdings nicht riesig. Bei allen anderen Variablen erweist sich die Web/Papier-Kombination als überlegen.

¹ Es gilt hier allerdings anzumerken, dass die Quotierung bezüglich des Stadtteiles nicht nur dafür verwendet wurde, dass jedes Quartier genügend Teilnehmende erreicht, um das richtige Verhältnis in der Population zu erhalten, sondern auch um einige Quartiere, die zu klein sind, auf das anvisierte Minimum zu bringen. Somit ist eine künstliche Überrepräsentation der kleinen Quartiere die Folge. Werden die kleinen Quartiere aus der Berechnung ausgeschlossen, verbessert sich natürlich die durchschnittliche Repräsentativität der CATI-Studie. Aber die Web/Papier-Kombination erzielt auch bei dieser Auswahl eine bessere Repräsentativität.

Für diese Analysen wurden immer die ungewichteten Daten verwendet. Häufig werden Gewichtungen verwendet, um Verzerrungen in der Verteilung zu korrigieren. Allerdings besteht auch hier das Problem, dass nur Variablen für die Gewichtung verwendet werden können, von welchen man die Verteilung in der Population kennt. Somit entsteht das Risiko, dass zwar mit der Gewichtung eine gute Repräsentativität bezüglich der für die Gewichtung verwendeten Variablen (im Normalfall sozio-demographische Variablen) erreicht, dafür aber in Bezug auf andere Variablen die Verzerrung vergrössert wird. Dann nämlich, wenn die Gewichtungsvariablen nicht mit der Zielvariable korrelieren, letztere aber verzerrt ist (oder in anderen Worten: die Teilnahme von der Zielvariable direkt oder indirekt abhängt). Nehmen wir zum Beispiel an, dass sich in der Population die Lebenszufriedenheit zwischen den Geschlechtern nicht unterscheidet. Zufriedene Menschen nehmen aber eher an Befragungen teil. Gleichzeitig nehmen mehr Frauen als Männer an der Untersuchung teil. Weil die Teilnahme von der Lebenszufriedenheit abhängt (in anderen Worten: in der Stichprobe eine Verzerrung bezüglich Lebenszufriedenheit besteht), ist in der Stichprobe die Lebenszufriedenheit der Männer höher als die der Frauen. Wird nun nach Geschlecht gewichtet, werden Männer zum Beispiel mit dem Faktor 1.1 hochgewichtet, um zur korrekten Verteilung des Geschlechts zu kommen. Somit werden die zufriedenen Männer noch stärker überrepräsentiert (weil Männer mehr Gewicht kriegen) und die Lebenszufriedenheit der Männer wird in der Stichprobe durch die Gewichtung zusätzlich erhöht.

In der CATI-Studie wurde immer bezüglich der Quotenzellen gewichtet (eine Gewichtung nach der Interaktion zwischen AusländerIn/SchweizerIn, Geschlecht und Stadtteil). Im Folgenden wird kurz auf den Effekt der Gewichtung auf die Variablen in der Population eingegangen (s. Tabelle 6).

Tabelle 6. Durchschnittliche prozentuale Abweichung des erreichten Anteils vom erwarteten Anteil teilnehmender Personen per Variable: Effekt der Gewichtung nach den Quotierungsvariablen auf andere sozio-demographische Variablen

	CATI	CATI gewichtet
Alter	6.1	6.7
Alter CH	7.2	7.5
Alter Ausländer	5.4	5.6
Nation	17.0	15.7
Bildung	4.2	3.3
Summe (Alter, Bildung, Nation)	27.3	25.7

Die Resultate zeigen, dass die Gewichtung insgesamt bezüglich der drei Variablen einen positiven Effekt auf die Repräsentativität hat. Allerdings wird aus Tabelle 6 auch ersichtlich, dass die Gewichtung auch zu einer Verstärkung der Verzerrung führen kann, denn die Gewichtung führt zu einer grösseren Verzerrung bezüglich des Alters. Der Effekt der Gewichtung auf die Verzerrung kann auch innerhalb einer Variablen unterschiedlich sein. So ist die Verbesserung der Repräsentation der Variable Nationalität beispielsweise fast ausschliesslich auf die Verbesserung der Repräsentation bezüglich Personen aus Deutschland und der Schweiz zurückzuführen (zwei der am stärksten verzerrten Gruppen), während sich die Repräsentation anderer Nationalitäten durch die Gewichtung verschlechtert, wie Tabelle 7 zeigt.

Tabelle 7 ist zusätzlich zu entnehmen, dass die schlechtere Repräsentation der Nationalitäten der Web/Papier-Kombination grösstenteils auf die massive Überrepräsentation der Schweizer und Schweizerinnen zurückzuführen ist.

Die Repräsentation der Nationalitäten gibt auch einen Aufschluss über den Effekt, den die verschiedenen Befragungssprachen auf die Repräsentation haben können. Die CATI-Studie wurde in sechs Sprachen angeboten, der Online- und der Papierfragebogen jedoch nur in Deutsch. Die massive Überrepräsentation der SchweizerInnen lässt vermuten, dass dies ein möglicher Effekt der Einsprachigkeit ist. Zusätzlich ist festzustellen, dass die Nationalitäten der Alternativsprachen Italienisch, Englisch, Portugiesisch, Spanisch und Serbokroatisch etwas besser repräsentiert sind in CATI als in der Web/Papier-Kombination. Allerdings macht dies mengenmässig relativ wenig aus.

Die Befragten wurden auch nach der Sprache gefragt, die sie am besten beherrschen. Hier zeigt sich ebenfalls lediglich ein starker Effekt bezüglich der Deutsch sprechenden Personen (die allerdings sowieso die grosse Mehrheit ausmachen): Während bei der CATI-Studie rund 76% Schweizerdeutsch oder Deutsch sprechen, sind es bei der Web/Papier-Kombination 86%.

Bezüglich der anderen Nationen ergibt sich, dass der Anteil an Englischsprachigen in beiden Stichproben etwa ähnlich ist, die spanisch, portugiesisch und serbokroatisch Sprechenden aber deutlich weniger vertreten sind in der Web/Papier-Umfrage, die italienisch Sprechenden etwas weniger deutlich. Ebenfalls unterverteten sind die französisch Sprechenden. Das Anbieten einer zusätzlichen Sprache kann also durchaus die Teilnahme dieser Personengruppe fördern.

Tabelle 7. Prozentuale Abweichung des erreichten Anteils vom erwarteten Anteil teilnehmender Personen per Nationalitätengruppe: CATI, gewichtetes CATI und Web/Papier

	CATI	CATI gewichtet	Web/Papier
Schweiz	-1.0	0.4	7.2
Deutschland	2.8	1.9	0.0
Italien	-3.4	-3.4	-3.2
Portugal	-0.3	-0.3	-0.4
Serbien, Montenegro und Kosovo	-0.6	-0.5	-0.9
Spanien	0.2	0.2	-0.5
Österreich	0.5	0.3	0.7
Frankreich	-0.2	-0.3	-0.5
Türkei	-0.3	-0.3	-0.4
Grossbritannien	-0.3	-0.3	-0.6
Indien	-0.3	-0.4	-0.5
USA	-0.2	-0.3	-0.2
Griechenland	0.1	0.1	-0.2
Mazedonien	-0.4	-0.4	0.0
China (Volksrepublik China)	-0.4	-0.4	-0.4
Polen	-0.1	-0.2	-0.2
Kroatien	0.0	0.1	-0.3
Europa	4.4	4.3	1.9
Asien	-0.3	-0.4	-0.8
Amerikas	0.3	0.4	-0.2
Afrika	-0.7	-0.8	-0.6
Ozeanien	0.1	0.1	0.2
Summe	17.0	15.7	20.0
Summe ohne CH und D	13.2	13.4	12.7
Summe ohne CH	15.9	15.3	12.8

Bemerkung. Die Summen werden durch die absoluten Beträge der Abweichungen berechnet, sodass sich positive und negative Abweichungen nicht gegenseitig aufheben.

4 Auswirkungen eines Mode-Wechsels auf die Verteilungen der Variablen

Die Analyse der Repräsentativität der verschiedenen Modi weist darauf hin, dass die Kombination Web/Papier deutliche Vorteile gegenüber dem Weiterführen der Befragung im CATI-Mode aufweist.

Ein Wechsel des Modus hat aber nicht nur Auswirkungen auf die Repräsentativität einer Studie. Ein solcher Wechsel kann sich auch auf die Verteilungen der Variablen und somit auf die Resultate der Studie auswirken. Dies wird in der Literatur als Mode-Effekt bezeichnet. Da die Zeitreihe der Bevölkerungsbefragung weitergeführt werden soll, interessiert der Unterschied zwischen den Ausprägungen der Variablen in der CATI-Studie und jenen der Web/Papier-Kombination, jener Alternative, die sich in der Analyse der Repräsentativität als am geeignetsten erwiesen hat.

Grundsätzlich unterscheidet man zwei Typen von Mode-Effekten basierend auf der Art, wie der Unterschied eines Resultates zwischen zwei Modes zustande kommt: Der Anteil des Mode-Effekts, der darauf zurückzuführen ist, dass die verschiedenen Modes unterschiedliche Personen ansprechen, wird als Selektions-Effekt bezeichnet. Jener Anteil, der darauf zurückzuführen ist, dass die gleichen Personen unter sonst gleichen Bedingungen je nach Mode (leicht) unterschiedliche Antworten geben, wird als Methoden-Effekt bezeichnet (manchmal auch Measurement-Effekt).

Selektionseffekte sind in der Theorie eng mit Repräsentation verbunden. In der Praxis ist dies nicht so einfach aufzuzeigen, denn die Repräsentativität kann nur für Variablen beurteilt werden, von denen die „wahre“ Verteilung in der Population bekannt ist. Deshalb sind Selektionseffekte und Repräsentativitätsfragen in der Praxis getrennt zu behandeln. Ein Beispiel soll dies verdeutlichen: Während mehr ältere Menschen einen Festnetzanschluss besitzen und deutlich häufiger an CATI-Studien teilnehmen als jüngere, ist es bei Web-Studien umgekehrt. Bezüglich der Repräsentation besteht hier also ein Unterschied. Nun könnte es zusätzlich noch sein, dass besonders jene älteren Leute einen Internet-Zugang besitzen, die weltoffen und aktiv sind. So würde neben dem Selektionseffekt, der durch klassische Repräsentativitätsanalysen entdeckt würde, noch ein weiterer Selektionseffekt hinzukommen, wenn beispielsweise Fragen über die Reisefähigkeit gestellt würden: Die älteren Menschen in der CATI-Studie würden deutlich weniger reisen als die älteren Menschen in der Web-Studie. Somit würde sich der Effekt des Alters auf die Reisetätigkeit in der CATI-Studie vom Effekt in der Web-Studie unterscheiden, auch wenn bezüglich der Repräsentativi-

tät (Anteil der älteren Menschen in der Stichprobe), beispielsweise durch Gewichtung, korrigiert würde.

Methodeneffekte bezeichnen Effekte, die durch unterschiedliches Antwortverhalten aufgrund der Methode entstehen. So geben zum Beispiel Teilnehmende eher sozial erwünschte Antworten, wenn ein Interviewer präsent ist (z.B. bei CATI oder persönlichen Interviews). Weitere Methodeneffekte sind etwa der sogenannte „recency-Effekt“ bei telefonischen und mündlichen Befragungen, der dadurch zustande kommt, dass Respondenten eher die letzten Antwortkategorien wählen, da diese nach einer Aufzählung leichter zu erinnern sind. Das Gegenstück ist der „primacy“-Effekt: In schriftlichen Erhebungsmodi wählen Respondenten eher die ersten Kategorien (oder die erst-beste Kategorie, was jedoch manchmal separat als „satisficing“ geführt wird).

Die Trennung zwischen Selektions- und Methodeneffekten ist sehr schwierig, da sie vermischt sind. Selbst mit komplexen Experimentaldesigns ist es nur teilweise möglich, abzuschätzen wie gross der Anteil des Selektionseffektes am gesamten Methodeneffekt ist (vgl. für komplexe Designs, die aber auf starken Annahmen basieren und auch nicht eindeutig trennscharf sind: Schouten et al., 2013; Vannieuwenhuyze, Loosveldt & Molenberghs, 2012).

Im Rahmen dieses Reports kann hier nicht auf die Trennung von Selektions- und Methodeneffekten eingegangen werden. Ebenfalls nicht berücksichtigt werden Unterschiede bezüglich „primacy“- und „recency“-Effekten. Stattdessen liegt der Fokus der Analyse auf dem Gesamteffekt eines Wechsels von CATI auf Web/Papier.

Im Folgenden werden zuerst allgemeine Erkenntnisse zur Vergleichbarkeit der CATI-Studie mit der Web/Papier-Kombination beschrieben. Anschliessend wird auf die Gewichtung eingegangen. Schliesslich wird anhand von einigen Beispielen aus dem Bericht der Bevölkerungsbefragung 2013 genauer dargelegt, wie sich ein Wechsel von CATI auf Web/Papier auswirken kann.

4.1 Allgemeine Erkenntnisse zur Vergleichbarkeit der CATI-Studie mit der Web/Papier-Kombination

Um den (Gesamt-)Effekt eines Modewechsels auf die einzelnen Variablen zu prüfen, eignet sich grundsätzlich ein t-Test. Der t-Test gibt an, ob sich die Mittelwerte einer Variablen zwischen zwei Modes statistisch signifikant unterscheiden. Neben der statistischen Signifikanz, die angibt, ob die Mittelwerte aus der Stichprobe auch in der Population unterschiedlich sind, ist allerdings auch die Grösse des Unterschiedes resp. die Effektstärke von Belang. Ein gut interpretierbares Mass für die Stärke eines Effektes ist das sogenannte „Cohen's d“ (Cohen, 1977), das bei t-Tests einfach berechnet werden kann. Cohen (1992)

erstellte die Daumenregel, dass Effektstärken von $d > 0.20$ einem kleinen Effekt entsprechen, $d > 0.50$ einem mittleren Effekt und $d > 0.80$ einem grossen Effekt. Dies ist allerdings nur eine Daumenregel und hängt von den Umständen ab (s. Durlak, 2009). Die Beurteilung der Effektstärke ist wichtig, da die statistische Signifikanz sehr stark von der Stichprobe abhängt. Für die Parallelstudie wurde eine relativ grosse Stichprobe gezogen, damit für die Punktschätzer für die Zeitreihe eine genügende statistische „Power“ erreicht wird. Deshalb werden beim Vergleich von Mittelwerten zwischen der CATI-Studie und der Web/Papier-Kombination schon relativ kleine Effektstärken als statistisch signifikant ausgewiesen ($N=3'672$). Ein statistisch signifikanter Unterschied bedeutet also nicht zwingend, dass sich der Unterschied in einem für die inhaltliche Interpretation relevanten Bereich bewegt.

Der t-Test setzt eigentlich ein kontinuierliches Skalenniveau voraus, während die meisten in der Bevölkerungsumfrage verwendeten Fragen ein ordinales Skalenniveau aufweisen (sechsstufige Notenskalen und vier- oder fünfstufige Likertskalen). Häufig werden allerdings auch für ordinale Likertskalen Analysemethoden verwendet, die eigentlich ein kontinuierliches oder gar intervallskaliertes Skalenniveau annehmen. Gerade Mittelwerte können bei schiefverteilten ordinalen Skalen jedoch ziemlich verzerrt sein. Deshalb wurde im Sinne einer Sensitivitätsanalyse zusätzlich für alle ordinalen Skalen ein ordinales logistisches Modell verwendet, um signifikante Unterschiede zwischen den Verteilungen der Variablen zwischen den Modes zu untersuchen und mit den Resultaten der t-Tests zu vergleichen (vgl. z.B. auch Martin & Lynn, 2011). Grundsätzlich sind die Resultate deckungsgleich. Lediglich dort, wo die Cohen's d sehr klein sind und auch das Signifikanzniveau beim t-Test knapp unter 0.05% liegt, zeigen sich Unterschiede dahingehend, dass der ordinale Test keinen signifikanten Unterschied zwischen den Verteilungen findet. Deshalb werden im Folgenden für einen allgemeinen Überblick die t-Tests und Cohens's d interpretiert, wann immer sowohl der t-Test wie auch das ordinale Modell einen signifikanten Unterschied anzeigen und das Cohen's d nicht unterhalb von 0.18 fällt (somit wird etwas konservativer vorgegangen als von Cohen empfohlen). Im nächsten Abschnitt wird dann anhand einiger Beispiele genauer auf Unterschiede in den Verteilungen eingegangen (und somit das Skalenniveau korrekt berücksichtigt). Der allgemeine Überblick in diesem Abschnitt bezieht sich ausserdem insbesondere auf ordinale, dichotome und intervallskalierte Variablen. Kategoriale Variablen und offene Fragen erfordern eine komplexere Analyse und werden deshalb hier nur mit einem Beispiel berücksichtigt. Der Fragebogen enthält allerdings auch nur wenige kategoriale Variablen.

Im Allgemeinen zeigen sich häufig die erwarteten Effekte: Die Teilnehmenden der Web/Papier-Kombination geben tendenziell kritischere Antworten. Dies hat

sicherlich zwei Ursachen: Der eine Grund liegt in Unterschieden der Selektion. An der CATI-Studie nehmen nur wenige Leute teil (Rücklaufquote von lediglich 33% im Vergleich zu 56% der Web/Papier-Kombination). Insofern werden eher Menschen zur Teilnahme überzeugt, die grundsätzlich zufriedener mit der Stadt sind. Meistens fehlen in Befragungen die unzufriedeneren Menschen (s. z.B. Martin & Lynn, 2011). Der zweite Grund liegt in den unterschiedlichen Methoden. Die CATI-Studie ist interviewerbasiert. Das heisst, Interviewer sprechen mit den Teilnehmenden. Dies führt dazu, dass diese eher positivere Antworten geben als wenn sie ohne eine gegenüberstehende Person eine Antwort geben würden (soziale Erwünschtheit, s. z.B. Bowling, 2005; Kreuter, Presser & Tourangeau, 2008; Martin & Lynn, 2011).

Konkret zeigt sich dies insbesondere bei den Variablen zur Zufriedenheit mit Angeboten der Stadt. Von den 14 Variablen zu verschiedenen Einrichtungen und Angeboten weist einzig die Zufriedenheit mit der Sauberkeit keinen statistisch signifikanten Unterschied zwischen der CATI-Studie und der Web/Papier-Kombination auf. Allerdings erreichen auch nur sechs Variablen eine Effektstärke von mehr als 0.20: Zufriedenheit mit der Kinderbetreuung, dem Wohnungsangebot, dem öffentlichen Verkehr, dem Parkplatzangebot, der Verkehrssicherheit und den Schulen. Den stärksten Effekt weist die Zufriedenheit mit dem Wohnungsangebot auf ($d=-0.40$; $p<0.001$).

Bezüglich der verschiedenen Massnahmen, die die Stadt Zürich ergreift (oder nicht), ergibt sich ein differenzierteres Bild. Die Effektstärken liegen mit einer Ausnahme (Schaffung von Arbeitsplätzen) alle unter der $d=0.20$ -Schwelle. Bezüglich der statistischen Signifikanz zeigt sich, wenn man sowohl die Signifikanzniveaus der ordinalen Regression wie auch der t-Tests zum Mass nimmt, dass sich nur sechs von 14 Variablen zwischen den beiden Modes unterscheiden (Massnahmen für die öffentliche Sicherheit, die Schaffung von Arbeitsplätzen, die Wirtschaftsförderung, gegen Drogenmissbrauch, für die Umweltsituation und für den öffentlichen Verkehr). Bei diesen Fragen ist zu bemerken, dass die Teilnehmenden der Web/Papier-Kombination eher als die Teilnehmenden der CATI-Studie finden, dass zu wenig unternommen wird.

Bezüglich der Fragen zur wirtschaftlichen Lage zeigen sich nur sehr kleine Unterschiede. Einzig die Beurteilung der eigenen aktuellen wirtschaftlichen Lage fällt bei der Web/Papier-Kombination etwas schlechter aus ($d=-0.32$; $p<0.001$). Kein Effekt ist bei der Beurteilung des Ausgabeverhaltens der Stadt Zürich sowie der Einschätzung der eigenen Wirtschaftslage in 1-2 Jahren festzustellen. Die Effekte der Einschätzung der allgemeinen aktuellen und zukünftigen (1-2 Jahre) Wirtschaftslage fallen vernachlässigbar gering aus ($d= -0.11$; $p<0.01$; resp. $d=0.09$; $p<0.01$).

Überraschenderweise unterscheiden sich die Resultate der CATI-Studie nicht von jenen der Web/Papier-Kombination bezüglich des Interessens für die Politik in der Stadt Zürich. Es lässt sich aber ein kleiner Unterschied bezüglich des Gefühls, wie gut man sich von den Behörden vertreten fühlt, feststellen ($d=-0.19$, $p<0.001$).

Die Zufriedenheit mit den verschiedenen Verwaltungsabteilungen der Stadt Zürich fällt in der Web/Papier-Kombination durchgängig tiefer aus als in der CATI-Kombination (Effektstärken von $d=-0.18$ bis $d=-0.22$, $p<0.001$).

Fast alle Fragen zur Wohnsituation weisen einen signifikanten Unterschied zwischen der CATI-Studie und der Web/Papier-Kombination auf. Lediglich die Höhe des Mietzinses und die Veränderung der Wohnumgebung zeigen keinen Mode-Effekt. Allerdings sind die Effektgrößen aller Variablen sehr klein. Bei lediglich drei Variablen liegt das Cohen's d über 0.18 (Empfindung der Höhe der Miete bezüglich dem Budget, $d=0.24$; Zufriedenheit mit der Wohnung, $d=-0.25$; Zufriedenheit mit der Wohnumgebung, $d=-0.20$). Die Effekte verlaufen auch hier in die erwartete Richtung: Weniger Zufriedenheit unter den Web/Papier-Teilnehmenden.

Alle Variablen zum Thema Quartier und zur Umzugsabsicht weisen signifikante Effekte auf, und zwar wiederum alle in die Richtung stärkerer Unzufriedenheit resp. höherer Umzugsabsicht unter den Teilnehmenden der Web/Papier-Kombination. Hier fallen auch die Effektstärken etwas grösser aus: sieben der elf Variablen weisen eine Effektstärke über $d=0.18$ auf, zwei über $d=0.30$ (Nachbarschaft, $d=-0.31$; Bevölkerungszusammensetzung, $d=-0.30$; öffentlicher Verkehr, $d=-0.24$; Sauberkeit, $d=-0.20$; Verkehrssicherheit, $d=-0.26$; öffentliche Sicherheit, $d=-0.26$; Ruhe, $d=-0.23$).

Bezüglich des Sicherheitsgefühls lassen sich kaum Effekte ausmachen. Nur die allgemeine Frage nach dem Sicherheitsgefühl zeigt einen signifikanten Effekt, die Effektstärke ist mit $d=-0.09$ aber sehr klein.

Die meisten Verkehrsfragen weisen einen signifikanten Effekt auf. Lediglich die Häufigkeit, mit welcher die Teilnehmenden mit dem Fahrrad oder dem öffentlichen Verkehr unterwegs sind, unterscheidet sich nicht zwischen den beiden Modes (die ordinale Regression findet jedoch für das Fahrrad einen leicht signifikanten Unterschied²). Die Effektstärken sind auch hier im kleinen Bereich, aber dennoch nicht zu vernachlässigen (Zufriedenheit als Velofahrer, $d=-0.24$; Zufriedenheit wenn mit dem öffentlichen Verkehr unterwegs, $d=-0.22$; Häufigkeit

² Dies weist darauf hin, dass der Durchschnitt sich nicht verändert hat (t-Test), aber durchaus in der Verteilung Unterschiede bestehen, siehe Abschnitt 4.3.

mit dem Auto unterwegs, $d=0.21$; Zufriedenheit als Autofahrer, $d=-0.30$; Kenntnis des Projektes „Stadtverkehr 2025“, $d=-0.18$).

Schliesslich zeigt sich ein signifikanter Effekt für die Frage, ob man die Studie sinnvoll findet. Wenig überraschend finden die Web/Papier-Teilnehmenden die Studie etwas weniger sinnvoll, die Effektstärke ist aber auch hier ziemlich gering ($d=-0.20$).

Überraschenderweise zeigen sich lediglich im t-Test signifikante, aber extrem kleine Effekte ($d=-0.07$ und -0.08) bei der Beurteilung, wie gerne die Teilnehmenden in der Stadt Zürich wohnen und bei der Beurteilung der Lebensqualität. Bei den ordinalen logistischen Modellen sind die Effekte sogar statistisch insignifikant.

Um Mode-Effekte bezüglich offener Fragen zu untersuchen, bietet sich die Frage nach den grössten Problemen der Stadt Zürich an. Hier zeigt sich, dass die offene Fragestellung der Web/Papier-Kombination einige Vorteile gegenüber der offenen Fragestellung in der CATI-Studie bietet: Bei der Web/Papier-Kombination werden die Daten so erfasst, wie sie von den Teilnehmenden eingegeben werden. Eine Kodierung für die Auswertung findet ex-post, resp. im Rahmen der Auswertung statt. Das heisst, dass erstens die Kodierung durch ausgewählte Kodierer erfolgt, im Optimalfall diejenigen Personen, die auch die Auswertung durchführen, und dass zweitens die Rohdaten auch in Zukunft zugänglich sind. Die eigene Kodierung reduziert den Effekt, den die Interviewer auf die Kodierung haben, und es kann die Qualität der Kodierung systematisch getestet werden (z.B. durch die Interrater-Reliabilität, wenn zwei oder drei Personen die Kodierung parallel durchführen). Bei der CATI-Studie können nämlich die Interviewer auf mindestens zwei Arten die Antworten stark beeinflussen: einerseits können sie durch die Kenntnis der vorgegebenen Kodierungsmöglichkeiten den Teilnehmenden bereits suggestiv eine Antwort vorschlagen, andererseits können sie die Antworten zu schnell und nach zu einfachem Schema kodieren. Die Verfügbarkeit der Rohdaten bei der Web/Papier-Kombination bringt den Vorteil mit sich, dass nachträgliche Änderungen im Kodierschema möglich werden: Zeigt sich in acht Jahren, dass plötzlich ein neues Problem relevant wird, lässt sich so überprüfen, wieviele Personen schon etwas in diese Richtung erwähnt haben (beispielsweise würden Probleme mit Elektrofahrrädern im Moment wohl unter „Verkehr“ kodiert. Würde nun ersichtlich, dass dies eine solche Dimension erreicht, dass eine eigene Kategorie nötig wird, liesse sich das in den Rohdaten rückverfolgen).

Der Effekt, den Interviewer auf die Kodierung haben können, spiegelt sich denn auch in den Daten der Studien dieser Methodenevaluation wieder: Die Antwortkategorien, welche die Interviewer als vorkodierte Antwortkategorien zur Verfü-

gung hatten³, werden in der CATI-Studie massiv häufiger verwendet als in der Web/Papier-Kombination. Lediglich die Arbeitslosigkeit wird gleich oft genannt und in der Web/Papier-Kombination wird etwas häufiger moniert, dass die Steuern zu hoch sind. Insgesamt machen die vorgegebenen Kategorien bei der CATI-Studie bis zu rund zwei Drittel der Antworten aus (1. Antwort: 66%, 2. Antwort: 59%, 3. Antwort: 44%), während bei der Web/Papier-Kombination andere Antworten deutlich häufiger sind (1. Antwort: 37%, 2. Antwort: 30%, 3. Antwort: 29%). Dennoch zeigt sich, dass die Web/Papier-Antworten genau so kodiert werden können wie die CATI-Antworten, denn die Befürchtung, dass die Teilnehmenden der Web/Papier-Kombination Antworten geben würden, die nicht in das Kodierschema passen oder nichts mit der Frage zu tun haben, hat sich als nicht zutreffend herausgestellt (die Kategorie „Anderes“ musste zwar etwas häufiger aber dennoch selten angewendet werden: bei CATI 1%-3%, bei Web/Papier 2%-4%).

Dass die Teilnehmenden in der Web/Papier-Kombination die offene Frage selbst ausfüllen, bringt allerdings auch Nachteile mit sich: Erstens muss die Kodierung ex-post stattfinden, was mit einem Aufwand verbunden ist, der in der CATI-Studie von den Interviewern „on the go“ erledigt wird. Der zusätzliche Aufwand ist aber, wie oben beschrieben, eher als Vorteil zu sehen denn als Nachteil, da dies eine Kontrolle und Auswertungsmöglichkeiten mit sich bringt. Ein zweiter Nachteil besteht im Risiko, dass die Teilnehmenden die Frage nicht ganz korrekt beantworten. Bei der Frage zu den drei grössten Problemen der Stadt könnte dies beispielsweise sein, dass die Teilnehmenden mehr als drei Probleme nennen. Die Daten der Methodenevaluation zeigen jedoch, dass auch in der CATI-Studie 4.4% der Befragten mehr als drei Probleme nannten. Dies ist erstaunlich, da die Interviewer eigentlich angewiesen sind, *drei* Probleme aufzunehmen. In der Web/Papier-Kombination fällt dieser Anteil allerdings grösser aus (11.3% resp. 12.6%). Dieses Problem ist aber leicht zu umgehen, da die drei ersten Nennungen ausgewählt werden und die anderen Nennungen nicht beachtet werden können (wie dies eigentlich auch in der CATI-Studie der Fall wäre, wenn Interviewer weitere Nennungen ignorieren).

Da dem Autor bei der CATI-Studie nur die vorkodierten Antworten zur Verfügung standen und da es bezüglich der drei Probleme Doppel- und Dreifachnennungen gab (beispielsweise „Verkehr“ oder „Baustellen“ als erstes, zweites und drittes Problem), konnte keine genauere Analyse der Unterschiede zwischen der CATI-Studie und der Web/Papier-Kombination durchgeführt werden.

³ Für diese Analyse wurde die vom Erhebungsinstitut vorkodierte Variable verwendet, die 89 Kategorien aufweist. Dabei wurde die Analyse auf die ersten drei Antworten eingeschränkt.

4.2 Auswirkungen der Gewichtung

Die Ausführungen im oberen Abschnitt beziehen sich auf die ungewichteten Daten, damit die beiden unterschiedlichen Erhebungsmodi direkt miteinander verglichen werden können. In den Publikationen zu den Bevölkerungsbefragungen wurde jedoch jeweils eine Gewichtung verwendet, die auf den Quotierungszellen beruht. Die Gewichtung wird so vorgenommen, dass die Verteilung bezüglich der 92 Quotenzellen (Stadtquartier, Geschlecht, AusländerIn/SchweizerIn) in der Stichprobe mit derjenigen in der Population übereinstimmt.

Wie im Abschnitt 3.2 aufgezeigt wurde, kann eine solche Gewichtung für einige Variablen zu einer Verringerung der Verzerrung führen. Es kann aber auch sein, dass die Gewichtung für einige Variablen zu einer Verstärkung der Verzerrung führt. Somit interessiert es natürlich auch, wie sich eine Gewichtung auf den Mode-Effekt auswirkt.

Wird für die CATI-Studie die oben beschriebene Gewichtung verwendet, zeigt sich, dass in fast sämtlichen Fällen der Mode-Effekt ganz leicht verringert wird oder gleich bleibt. Eine Vergrößerung des Unterschiedes findet sich nur bei 10 von 77 untersuchten Variablen. Dabei liegt die Vergrößerung des Effektes meist lediglich bei $d=0.01$, maximal bei $d=0.02$. Wird für die WebPapier-Kombination ebenfalls eine Gewichtung nach gleichem Prinzip vorgenommen, sind die Veränderungen bezüglich des Effektes unterschiedlicher: Einige verringern sich noch weiter, andere werden leicht grösser, die Veränderungen liegen auch hier lediglich im Bereich von $d=0.01$ bis 0.04 . So bleiben auch hier die Hauptresultate bestehen: Bezüglich der Schwellenwerte des Cohen's d ändert sich nichts substantiell durch die Gewichtung.

4.3 Konkrete Erläuterungen an Beispielen aus dem Bericht zur Bevölkerungsbefragung 2013

Zwei der zentralen Fragen in der Bevölkerungsbefragung sind die Wohnortbeliebtheit und die Lebensqualität der Stadt Zürich. Bei beiden Variablen zeigt sich eine stabil hohe Bewertung resp. eine leichte Steigerung. Diese setzt sich im aktuellen Jahr fort. Wie bereits der t-Test und das ordinale Modell gezeigt haben, besteht auch im Hinblick auf die Verteilungen kein relevanter Unterschied zwischen der Bewertung dieser beiden Variablen im CATI oder im Web/Papier-Mode.

Interessanter wird es, wenn Variablen in den Blick genommen werden, die im t-Test einen signifikanten Unterschied zwischen CATI und Web/Papier aufweisen. Als Beispiel können hier die Variablen zur Zufriedenheit mit den Einrichtungen und Angeboten verwendet werden. Es zeigt sich, dass zwar teilweise relativ grosse Unterschiede zwischen CATI und der Web-Kombination beste-

hen, diese sich aber kaum in einer sinnvollen Interpretation bemerkbar machen würden. Interessant ist, dass zwar diejenigen Variablen mit einem Cohen's d, das über dem Schwellenwert liegt, auch zu jenen Variablen gehören, bei denen der Punktschätzer der Web/Papier-Kombination tiefer liegt als die Werte seit 2005; allerdings gibt es auch Variablen mit einem Cohen's d über dem Schwellenwert, bei denen die Web/Papier-Kombination zwar tiefer als bei CATI liegt, aber durchaus im Rahmen der letzten Jahre. Werden die 95%-Vertrauensintervalle angeschaut, zeigt sich, dass die Vertrauensintervalle der WebPapier-Kombination nur bei wenigen Variablen nicht den Bereich der Werte früherer Jahre beinhalten (Kultur, Ausgangsmöglichkeiten, Wohnungsangebot, öffentlicher Verkehr und Parkplatzangebot). Die Werte der Web/Papier-Kombination sind mit Ausnahme der Sauberkeit überall tiefer als bei CATI. Dennoch bleibt im Vergleich der verschiedenen Einrichtungen und Angebote alles beim Gleichen: Die Unterschiede zwischen den Modes sind kleiner als die Unterschiede zwischen den Variablen. Bei einem Ranking ändern sich nur wenige Plätze und lediglich dort, wo sich die Bewertungen sowieso im gleichen Rahmen bewegen.

Als Beispiel für einen signifikanten t-Test mit jedoch kleinem Cohen's d (-0.09) dient die Frage „Sicherheitsgefühl in der Nacht“. Hier stieg in den letzten Jahren der Anteil der Personen, die sich sehr sicher oder eher sicher fühlen, kontinuierlich von 72% im 2005 über 73% im 2007, 75% im 2009 und 80% im 2011 auf 81% im 2013. Im 2015 ist dieser Wert weiter gestiegen und liegt in der CATI-Studie bei 85%. Obwohl der Wert in der Web/Papier-Kombination statistisch signifikant tiefer liegt, kommt er dennoch auf 84% zu liegen und ist somit auch deutlich höher als in den vorhergehenden Jahren. Hier zeigt sich wie bereits bei den Fragen zu den Einrichtungen und Angeboten, dass statistisch signifikante Unterschiede zwischen den Erhebungsmodi nicht zwingend substantiell von Bedeutung sind, insbesondere wenn die Effektstärke klein ist.

Interessant sind die Fragen zur Häufigkeit der Benutzung verschiedener Verkehrsmittel. Obwohl die Fragen in gleicher Form gestellt wurden, zeigt sich ein deutlicher Unterschied zwischen den Mode-Effekten. Während bei der Benutzung des Fahrrads ein deutlicher Rückgang jener zu verzeichnen ist, die nie mit dem Fahrrad unterwegs sind (2009: 58%; 2013: 55%; 2015 (CATI): 49%), ist der Anteil der Personen, die nie mit dem Auto (2009: 43%; 2013: 43%; 2015 (CATI): 45%) oder täglich mit den öffentlichen Verkehrsmitteln (2009: 46%; 2013: 46%; 2015 (CATI): 45%) unterwegs sind, sehr konstant. Der entsprechende Anteil der Web/Papier-Kombination liegt nun interessanterweise beim Fahrrad klar unter dem Anteil der CATI-Studie im gleichen Jahr (40% vs. 49%) und führt den sinkenden Trend verstärkt weiter. Bei den beiden konstant bleibenden Verkehrsmitteln zeigt sich bezüglich des Autos ein deutlich tieferer An-

teil an Personen, die nie das Auto benutzen (28% vs. 45%), während der Anteil der Personen, die täglich den öffentlichen Verkehr benutzen, praktisch gleich ist wie bei der CATI-Studie (44% vs. 45%).

Als Beispiel für eine offene Frage wird wiederum die Frage nach den drei grössten Problemen der Stadt Zürich verwendet. Die vom Erhebungsinstitut angebotene zusammenfassende Variable⁴ geht eher unkonventionell mit Mehrfachnennungen um: Es gibt einen nicht vernachlässigbaren Anteil an Personen, die zwei- oder dreimal dasselbe Problem nennen. Die Nennungen werden behandelt, als ob sie von verschiedenen Personen kommen würden. Dennoch wird hier zur Veranschaulichung diese Variable verwendet, was zu teils widersprüchlichen Resultaten zur Analyse der ausführlicher kodierten Variable der Problemnennungen führt. Die grundsätzliche Erkenntnis bleibt aber dieselbe: Der Effekt von dem Moduswechsel ist vernachlässigbar gering. Zwar unterscheiden sich einige Prozentzahlen zwischen CATI und der Web/Papier-Kombination, die Rangfolge bleibt aber mehr oder weniger ähnlich (Verschiebung von einem bis maximal zwei Rangplätzen). Somit sind die Unterschiede zwischen den Jahren grösser als zwischen den Erhebungsmodi. Einzige Ausnahme sind die Steuern und die Umweltbelastung: diese tauschen unter sich je nach Erhebungsmodus die Ränge 6 und 11 ab. Insbesondere bei jenen Problemen, die sich über der 5%-Marke bewegen (und es auch nennenswerte Unterschiede gibt), ändern sich jedoch die Rangfolgen zwischen CATI und der Web/Papier-Kombination gar nicht, obwohl bei den Wohnungsproblemen ein durchaus nennenswerter Unterschied in der Häufigkeit zu bemerken ist (s. Abbildung 2).

Zusammenfassend kann festgehalten werden, dass hinsichtlich der offenen Fragen bei der CATI-Studie ein deutlicher Interviewereffekt wirkt. Dieser kann durch die Web/Papier-Kombination aufgehoben werden. Die möglichen negativen Effekte offener Fragen bei der Web/Papier-Kombination erwiesen sich als vernachlässigbar gering. Im Allgemeinen sind bei einer groben Analyse der Probleme kaum Unterschiede festzustellen. Es muss aber berücksichtigt werden, dass sich diese Analyse auf lediglich eine offene Frage beschränkt.

Bezüglich der ordinal- und intervallskalierten Variablen kann keine allgemeingültige Aussage über den Effekt eines Modewechsels gemacht werden. Bei einigen Variablen ist ein Effekt sichtbar, bei anderen hingegen nicht. Dies ist nicht abhängig vom Skalen- oder Fragentyp. Hingegen sind einige Thematiken eher anfällig für einen Effekt: Die Zufriedenheit mit den Verwaltungsabteilungen der Stadt Zürich zum Beispiel ist in der Web/Papier-Kombination tiefer als in der CATI-Befragung.

⁴ Für diese Analyse wurde die vom Erhebungsinstitut vorkodierte Variable verwendet, die 13 Antwortkategorien anbietet. Wie im Bericht wurde die zusätzliche Kategorie „Anderes“ nicht berücksichtigt.

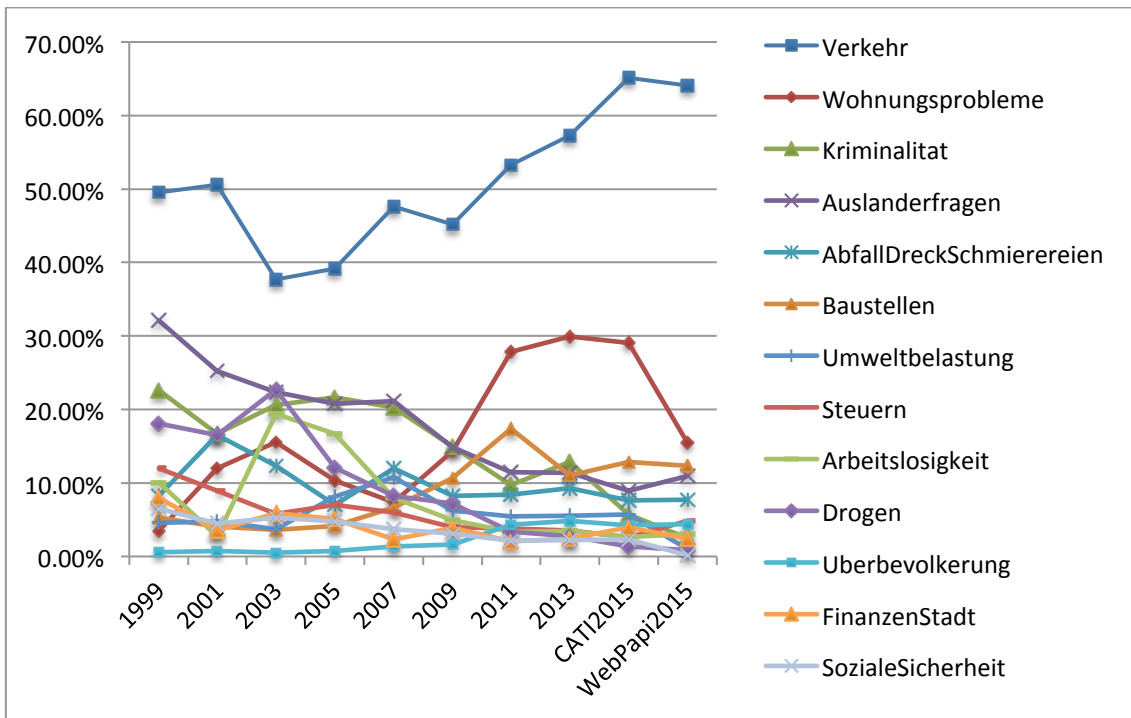


Abbildung 2. Grösste Probleme Zürichs. Veränderung über Jahre im Vergleich mit dem Moduswechsel von CATI zur WebPapier-Kombination im Jahr 2015

Grundsätzlich lässt sich festhalten, dass Fragen zur Zufriedenheit anfällig für einen Effekt sind (jedoch tritt nicht in jedem Fall ein Effekt auf). Dabei fallen die Werte in der Web/Papier-Kombination etwas kritischer aus: Die Befragten geben eher an, unzufriedener zu sein. Allerdings sind die Effekte meist zu klein, um einen inhaltlich relevanten Unterschied zu machen. Dies stimmt mit Ergebnissen anderer Mixed-Mode Studien überein (vgl. z.B. Martin & Lynn, 2011, S. 29).

5 Empfehlungen

Aus diesen Erkenntnissen der systematischen Methodenevaluation lassen sich folgende Empfehlungen formulieren bezüglich der Weiterführung der Bevölkerungsbefragung der Stadtentwicklung Zürich.

5.1 Umstellung des Erhebungsmodus auf Web/Papier

Die Untersuchung der Repräsentativität bezüglich der Stichprobe zeigte, dass die Web/Papier-Kombination den anderen getesteten Erhebungsmodi deutlich überlegen ist. Sogar bezüglich der Population ist die Web/Papier-Kombination basierend auf einer Zufallsstichprobe repräsentativer als die CATI-Erhebung basierend auf einer geschichteten Stichprobe mit Quotierung. Dies mag überraschen, hat aber damit zu tun, dass die Erreichbarkeit per Telefon je länger je schwieriger wird, zumal die Bewohner der Stadt Zürich oft per Telefon für mehr oder weniger seriöse Marktforschungsumfragen kontaktiert werden.

5.2 Zufallsstichprobe als Basis der Erhebung

Wie in der Parallelstudie sollte eine zukünftige Web/Papier-Erhebung auf einer reinen Zufallsstichprobe beruhen. Die Methodenevaluation hat gezeigt, dass eine Schichtung oder Quotierung nicht nötig ist. Bereits in einer in etwa halb so grossen Stichprobe, wie sie bisher für die Bevölkerungsumfrage verwendet wurde, sind die Verzerrungen in den einzelnen Quotenzellen, die in der Vergangenheit verwendet wurden, kleiner als in der CATI-Studie.

Die Quotierung auf Quartierebene wurde allerdings in der CATI-Studie nicht (nur) dazu verwendet, die Repräsentativität zu gewährleisten, sondern auch dazu, dass genügend Menschen aus den bevölkerungsschwächeren Quartieren den Fragebogen ausfüllen (rund 100 Teilnehmende pro Quartier). Dies soll auch Analysen auf der Quartierebene ermöglichen. Wird nun in einer Web/Papier-Erhebung auf eine Quotierung verzichtet, werden in kleineren Quartieren deutlich weniger Personen teilnehmen als in grösseren Quartieren, da die Methodenevaluation gezeigt hat, dass keine grossen Verzerrungen bezüglich der Quartiere zu erwarten sind. In der Web/Papier-Parallelstudie dieser Methodenevaluation mit einer Teilnehmerzahl von 1'171 Befragten variieren die Teilnehmerzahlen der Quartiere zwischen 23 und 89, mit einem Durchschnitt von 51 Personen. Um in einer allfälligen künftigen reinen Web/Papier-Befragung wie bisher auf die Gesamtteilnehmerzahl von mindestens 2'500 Personen zu kommen, wird die Stichprobe im Vergleich zur Parallelstudie mit dem Faktor 2.5 multipliziert werden müssen. Bei einer allfällig etwas geringeren Rücklaufquote würde dies bedeuten, dass der Durchschnitt bei rund 120 Personen liegen würde. Allerdings wären sieben Quartiere (Friesenberg, Hard, Kreis

5, Oberstrass, Fluntern, Witikon und Hirzenbach) mit weniger als 100 Personen vertreten (sechs davon deutlich). Jedoch sind auch in der CATI-Studie von 2015 sieben Quartiere mit weniger als 100 Personen vertreten (fünf davon mit 84-95 Personen). Da zusätzliche Interviews bei Web/Papier-Befragungen nur mit bescheidenem zusätzlichem Aufwand verbunden sind (marginale Grenzkosten), empfehlen wir, die Gesamtstichprobe zu vergrössern, sodass auch in den kleinen Quartieren eine hohe Teilnehmerzahl erreicht werden kann (z.B. durch eine dreifach so grosse Stichprobe wie in der Parallelstudie 2015). So wird weder eine Quotierung noch eine Gewichtung nötig, was eine hohe Repräsentativität der Resultate bezüglich aller Variablen ermöglicht. Eine zusätzliche Reservestichprobe (ebenfalls eine reine Zufallsstichprobe der Stadt Zürich) kann gezogen werden, falls Befürchtungen bestehen, dass in einigen Quartieren zu wenige Personen für eine Teilnahme motiviert werden können.

Müssen auf Quartierebene Aussagen über Unterschiede zwischen ausländischen und schweizerischen Personen sowie zwischen Männern und Frauen gemacht werden, wird es aufgrund des tiefen Ausländeranteils in einigen Quartieren schwierig, auf eine geeignete Anzahl ausländischer Frauen oder ausländischer Männer zu kommen. In diesem Fall gibt es, je nach Ziel der Untersuchung, verschiedene Möglichkeiten des Stichprobendesigns (wie z.B. eine zusätzliche Stichprobe für einzelne Quartiere, eine grössere Stichprobe in bevölkerungsschwachen Quartieren mit anschliessender sog. Designgewichtung). Es lohnt sich, im Vorfeld der Studie genau abzuklären, was analysiert werden soll, und dementsprechend ein Stichprobendesign zu wählen. Für eine optimale Implementierung der Stichprobenziehung(en) empfiehlt sich eine wissenschaftliche Beratung im Vorfeld der Studie, da sich die Zusammensetzung der Quartierbevölkerung bezüglich der interessierenden Variablen in Zwischenjahren ohne Befragung verändern kann. Es gilt allerdings zu berücksichtigen, dass eine allgemeine Bevölkerungsumfrage für eine detaillierte Analyse bezüglich kleiner Minderheiten grundsätzlich nicht geeignet ist.

5.3 Verzicht auf Gewichtung der Web/Papier-Kombination

Unsere Analysen haben gezeigt, dass eine Gewichtung nach den Variablen, wie sie in der Vergangenheit verwendet wurden (Quotierungsvariablen), zwar bezüglich einiger Variablen zu einer besseren Repräsentativität führt, bezüglich anderer Variablen aber auch Verzerrungen verstärken kann.

Bei der Analyse des Effektes eines Wechsels des Erhebungsmodus zeigte sich zudem, dass eine Gewichtung der Web/Papier-Kombination nicht nötig ist, weil sich die Gewichtung kaum auf die Resultate auswirkte.

Da mit der Web/Papier-Kombination bereits ohne Gewichtung und Quotierung die gleiche Repräsentativität wie in der CATI-Studie erreicht werden kann, sollte auf eine Gewichtung gänzlich verzichtet werden, um das Risiko einer Verstärkung der Verzerrung bezüglich einiger Variablen zu vermeiden. Ausserdem ist der Aufwand geringer und statistische Analysemethoden sind aussagekräftiger, wenn die Annahme gleicher Auswahlwahrscheinlichkeit der Respondenten nicht zu stark strapaziert wird. Schliesslich sind bei Gewichtungen einige (komplexere) statistische Methoden nicht mehr ohne weiteres anwendbar.

5.4 Weiterführung der Zeitreihe mit zwei Werten für 2015

Die Analyse der Effekte eines Wechsels des Erhebungsmodus auf die Antwortverteilungen hat gezeigt, dass diese sich im Hinblick auf eine substantielle Interpretation der Resultate in vernachlässigbaren Grössenordnungen bewegen. Dennoch bestehen teilweise Unterschiede zwischen den Erhebungsmodi CATI und WebPapier, die grösser sind als die Unterschiede über die Zeit (nämlich immer dann, wenn die Unterschiede über die Zeit sehr klein sind). Es ist schwierig, mit solchen Effekten umzugehen, weil erstens Effekte des Erhebungsmodus nicht bei allen Variablen gleich sind, es zweitens schwierig ist, solche Effekte genau vorherzusagen, und es drittens keine vordefinierten Lösungen oder Prozeduren gibt (wie etwa Gewichtung), wie Resultate für solche Effekte korrigiert werden könnten (vgl. Martin & Lynn, 2011).

Das Design der vorliegenden Methodenevaluation hat dieses Problem vorweggenommen und offeriert mit der Parallelstudie eine relativ einfache Lösung: Um den Lesenden der künftigen Berichte eine gut nachvollziehbare Zeitreihe bieten zu können, empfiehlt es sich, für die Werte von 2015 zwei Ausprägungen anzubieten: Einen für die CATI-Studie und einen für die Web/Papier-Kombination. So wird der Unterschied der Methoden im Zeitverlauf transparent bei gleichzeitiger Darstellung einer ungebrochenen Zeitreihe seit 1999. Als Beispiel für eine solche Zeitreihe (mit einer fiktiven Zahl für 2019) dient die Zufriedenheit mit dem Wohnungsangebot, die einen der stärksten Effekte aufwies: 2005: 3.53; 2007: 3.84; 2009: 3.23; 2011: 3.09; 2013: 3.08; 2015 (CATI): 3.26; 2015 (WebPapier): 2.75; 2019: 2.93.

5.5 Verwendung von 95%-Vertrauensintervallen in Graphiken

Damit die Lesenden die Grössenordnung der Unterschiede in der Zeitreihe sowie zwischen den Erhebungsmodi abschätzen können, sollten in den Graphiken die abgebildeten Zahlen mit 95%-Vertrauensintervallen versehen werden. So zeigt sich auch, ob sich die Unterschiede zwischen den Jahren in einem bedeutenden oder weniger bedeutenden Bereich bewegen. Die Weiterführung der Zeitreihe mit anderem Erhebungsmodus verstärkt diese Notwendigkeit, denn

die WebPapier-Kombination wurde im Jahr 2015, in welchem Daten für beide Erhebungsmodi vorliegen, mit einer halb so grossen Stichprobe erhoben wie die Daten der anderen Jahre. Dies vergrössert den Standardfehler teilweise beträchtlich.

5.6 Gründliche Evaluation der Zusammenarbeit mit einem Erhebungsinstitut resp. eigenständige Durchführung

Eine Befragung im Web/Papier-Modus ist im Vergleich zu einer CATI-Studie weniger aufwendig. Es lohnt sich zu prüfen, ob eine zukünftige Web/Papier-Studie gänzlich im Haus der Stadtentwicklung Zürich durchgeführt werden soll. Dies böte den Vorteil, dass alle Schritte unter Kontrolle der Stadtentwicklung Zürich sind und dass die Datenaufbereitung ebenfalls im Hause stattfinden kann. Somit kann der ganze Studienverlauf kontrolliert werden und es entstehen keine oder nur sehr geringe cash out-of-pocket Kosten.

Gängige Software zur Durchführung von Web/Papier-Befragungen bietet im Allgemeinen schon eine relativ gute Datenqualität für die Online-Befragung. Unterschiede bestehen im Erfassen der Papierfragebögen. Hier lohnt es sich im Vorfeld zu klären, ob die Papierfragebögen maschinenlesbar sind oder ob eine Software verfügbar ist, um eingescannte Fragebögen automatisch zu erfassen.

Freilich kann bei eigenständiger Durchführung keine telefonische Erinnerung realisiert werden. Es gilt deshalb zu prüfen, ob man bei geringem Rücklauf auch nach mehreren schriftlichen Erinnerungen ein Umfrageinstitut mit der telefonischen Erinnerung beauftragen möchte, was immer noch wesentlich billiger wäre als eine reine CATI-Erhebung.

Die eigenständige Durchführung hat den grossen Vorteil, dass kein kommerzielles Befragungsinstitut mit der Befragung verbunden ist. Gerade in den letzten Jahren hat die Belastung der Bevölkerung durch Marktforschungsumfragen stark zugenommen. Die grösseren Befragungsinstitute sind bekannt und nicht unbedingt beliebt. Eine Umfrage der Stadt Zürich hat sicherlich einen seriöseren Charakter und könnte auf weniger Ablehnung stossen, wenn die Befragung nicht mit dem Namen eines kommerziellen Befragungsinstituts verknüpft ist. Hingegen sind allfällige fachliche Beratungsdienstleistungen durch aussenstehende ExpertInnen nicht direkt ersichtlich und deshalb unproblematisch.

5.7 Zeitpunkt des Wechsels des Erhebungsmodus 2019

Wenn in einer Befragung Zeitreihen vorhanden sind, die weitergeführt respektive in regelmässigen oder unregelmässigen Abständen für Vergleiche herangezogen werden sollen, sind Änderungen im Befragungsdesign auf ein Minimum zu beschränken und bedacht durchzuführen. Fallen mehrere Änderungen an,

sollten diese möglichst zeitgleich stattfinden, damit es nur einmal zum Bruch kommt.

Einige Bedingungen müssen gegeben sein, damit ein Bruch in der Zeitreihe nicht dazu führt, dass die Zeitreihe gänzlich verloren geht: Erstens darf es nicht wiederholt zu Änderungen im Design kommen. Ansonsten können die Werte nicht mehr verglichen werden, da eine Referenz fehlt. Zweitens sollte abschätzbar sein, welche Konsequenzen ein Bruch auf die Zeitreihe hat. Drittens ist es von Vorteil, wenn die neue Methode mit einer Parallelstudie eingeführt wird.

Stadtentwicklung Zürich hat sich entschieden, die Punkte zwei und drei zu befolgen und hat mit der vorliegenden Parallelstudie Daten erhoben, die es erlauben, den Effekt eines Wechsels des Erhebungsmodus abzuschätzen, und die gleichzeitig – aufgrund der relativ grossen Stichprobe der Parallelstudie – auch die Zeitreihe im neuen Modus einführt, während der alte Modus in der gleichen Weise wie bisher realisiert wurde. Somit sind optimale Voraussetzungen gegeben, um den Wechsel des Erhebungsmodus zu begleiten. Deshalb ist es wichtig, dass ein Wechsel auch gleich bei der nächsten Erhebungsrunde vorgenommen wird, um die neue Zeitreihe nahtlos weiterzuführen.

Gleichzeitig wurde entschieden, dass die Bevölkerungsbefragung nicht mehr zweijährlich, sondern nur noch alle vier Jahre erhoben wird. Dies ist eine nicht so bedeutende Änderung des Designs wie die des Wechsels des Erhebungsmodus, da die Abschätzung des Effekts relativ einfach berechnet werden kann: man vergleicht die Zeitreihe in Vierjahresabständen und lässt bei den Analysen jeweils eine Erhebungsrunde aus. Es gibt allerdings noch weitere Effekte, die wegen eines solchen Designwechsels einen Einfluss haben könnten, die aber nicht abschätzbar sind: So könnte es beispielsweise sein, dass die Befragung an Bekanntheit einbüsst, wenn nur noch alle vier Jahre über sie informiert wird.

5.8 Pragmatisches, aber breites Sprachenangebot

Die CATI-Studie wurde bisher in sechs Sprachen durchgeführt, was mit einem erheblichen Aufwand verbunden ist. Die Parallelstudie wurde nur auf Deutsch realisiert. Zwar wurde eine relativ gute Repräsentation der Bevölkerung erreicht, allerdings zeigte sich auch eine Verzerrung bezüglich der Nationalität: an der Web/Papier-Kombination nahmen deutlich überproportional SchweizerInnen teil (rund 10% Überrepräsentation). Am meisten verwendet wurden bei der CATI-Studie die Sprachen Italienisch und Englisch (rund 7% der Interviews, die anderen vier Sprachen machten rund 3% aus). Bezüglich der Repräsentation waren aber Spanisch, Serbokroatisch und Portugiesisch wichtiger (obwohl sich die Unterschiede im 1%-Bereich bewegen). Da der Aufwand, einen Web/Papier-Fragebogen in verschiedenen Sprachen anzubieten, deutlich geringer ist als

derjenige, muttersprachliche TelefoninterviewerInnen zu finden und zu schulen, steht einem breiten Sprachenangebot grundsätzlich nichts im Wege.

Ein Effekt des Sprachenangebots für die Repräsentativität scheint aufgrund der Resultate dieser Studie sehr wahrscheinlich. Ob aber aufgrund der relativ kleinen Anzahl der fremdsprachigen TeilnehmerInnen ein Unterschied bezüglich der Resultate entsteht (also ob das Weglassen der zusätzlichen Sprachen zu einer inhaltlich relevanten Veränderung der Resultate führt), ist zumindest fraglich, sind doch die Unterschiede der Antworten zwischen CATI und Web/Papier sehr klein.

Insofern sollte das Sprachangebot pragmatisch definiert werden: genügend grosse Zielpopulation der Sprache, politisch wünschbare Subgruppe, notorisch unterrepräsentierte Subgruppen, finanzielle Überlegungen.

Die Wechsel des Modus von CATI auf Web/Papier einerseits und vom Zweijahres- auf den Vierjahresrhythmus andererseits sind kosteneffizient. Angesichts der Tatsache, dass die Mobilität und damit die (sprachliche) Heterogenität der Bevölkerung zunimmt, empfiehlt sich eine Investition eines gewissen Anteils der Einsparungen in zusätzliche Sprachen. Die Wahl der sechs bisherigen Sprachen erwies sich als effektiv, eventuell würde sich Französisch als eine weitere zusätzliche Sprache anbieten, handelt es sich dabei doch nicht nur um eine Landessprache, sondern auch um eine Sprache, die in vielen Ländern gesprochen wird und dementsprechend ein grosses Potential einer positiven Auswirkung auf die Repräsentativität mit sich bringt.

Die Qualität des Sprachenangebots sollte allerdings der Quantität vorgehen. Andernfalls leidet die Datenqualität. Die Übersetzung muss mindestens durch professionelle ÜbersetzerInnen erfolgen, Gold-Standard ist derzeit jedoch das TRAPD-Vorgehen (siehe für eine Erklärung: <http://ccsg.isr.umich.edu/translation.cfm>).

Appendix

Literatur

- Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised edition). New York, NY: Routledge.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.
- Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology 34*(9), 917-928.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646-675.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. *Public Opinion Quarterly, 72*(5), 847-865.
- Lipps, O. & Ochsner, M. (2016, forthcoming). Does providing additional languages improve representativeness? In T. P. Johnson, B. E. Pennell, I. Stoop, & B. Dorer (Hrsg.), *Advances in comparative survey methodology* (Wiley Series in Survey Methodology). New York, NY: Wiley.
- Martin, P., & Lynn, P. (2011). The effects of mixed mode survey designs on simple and complex analyses. *Centre for Coparative Social Surveys Working Paper Series, Paper No. 4, Nov. 2011.*
- Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research, 42*(6), 1555-1570.
- Stadtentwicklung Zürich. (2013). Bevölkerungsbefragung 2013. https://www.stadt-zuerich.ch/content/dam/stzh/prd/Deutsch/Stadtentwicklung/Publicationen_und_Broschueren/Stadt-_und_Quartierentwicklung/Befragungen/Bevoelkerungsbefragung_2013.pdf
- Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2012). A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. *International Statistical Review, 80*(2), 306-322.

Danksagung

Der Autor möchte sich bei Oliver Lipps, Michèle Ernst Staehli, Nicolas Pekari, Peter Farago und Rosa Sánchez Tomé für den Austausch über ihre Erfahrungen mit Mixed-Mode Experimenten insbesondere während der Design-Phase bedanken.