Methodische Probleme der Kausalanalyse bei sozialwissenschaftlichen Experimenten

Möglichkeiten zur Berücksichtigung von Effektheterogenität mithilfe von Kontextinformationen

Inaugural-Dissertation
zur Erlangung des Doktorgrades an der Sozialwissenschaftlichen Fakultät
der Ludwig-Maximilians-Universität München

vorgelegt von
Fabian Thiel

2024

Erstgutachterin:	Prof. Dr. Katrin Auspurg (Ludwig-Maximilians-Universität München)
Zweitgutachter:	Prof. Dr. Thomas Hinz (Universität Konstanz)
Tag der mündlichen Prüfung:	24.07.2023

Danksagung

Das Schreiben einer Dissertation ist sowohl fachlich als auch persönlich ein äußerst bereicherndes, zuweilen aber auch ebenso forderndes und langwieriges Unterfangen. Daher möchte ich mich bei allen Personen bedanken, die direkt oder indirekt zum Entstehen dieser Arbeit beigetragen und mich auf diesem Weg begleitet haben.

Mein herzlicher Dank gilt zunächst meiner Erstbetreuerin Prof. Dr. Katrin Auspurg für ihre wertvollen Anregungen, ihr konstruktives Feedback und ihre beständige Unterstützung. Sie gab den Ansporn, Ideen stets sorgfältig zu durchdenken und unermüdlich zu überarbeiten und weiterzuentwickeln. Meinem Zweitbetreuer Prof. Dr. Thomas Hinz möchte ich für den vertrauensvollen Spagat aus großem Freiraum in der Ausarbeitung und seinen bei Bedarf doch stets verfügbaren, hilfreichen Rückmeldungen danken. Für die Übernahme der Rolle der Drittprüferin in meiner Disputation bin ich Prof. Dr. Henrike Rau zu bestem Dank verpflichtet.

Zudem möchte ich Prof. Dr. Josef Brüderl und meinen (ehemaligen) Kolleg:innen für die vielen wertvollen Hinweise und Anmerkungen bei den verschiedenen lehrbereichsübergreifend organisierten Kolloquien und Workshops danken. Besonders möchte ich mich bei Sabine Düval, Werner Fröhlich, Christian Ganser und Andreas Schneck bedanken. Neben Euren hilfreichen Tipps und ehrlicher Rückmeldung habt Ihr auch stets für eine motivierende Arbeitsatmosphäre gesorgt und geschaut, dass mir nicht die Puste ausgeht.

Nicht zuletzt gilt meiner Familie ein herzlicher Dank für die Unterstützung und das Verständnis all die Jahre auf diesem Weg. Und schließlich möchte ich Dir, Elisabeth, ganz besonders herzlich danken. Für alles. Ohne Dich an meiner Seite hätte ich das sicher nicht geschafft.

Gröbenzell, November 2023 Fabian Thiel

Zusammenfassung

Experimente blieben in der Soziologie lange Zeit eher randständig, kommen in den letzten Jahrzehnten allerdings zunehmend zum Einsatz. Aufgrund ihrer hohen internen Validität gelten sie gemeinhin als „Goldstandard“ wissenschaftlicher Erkenntnis. Gerade Laborexperimente erlauben dabei ein besonders hohes Maß an Kontrolle über die Erhebungssituation, sehen sich allerdings oftmals auch dem Einwand ausgesetzt, dass aufgrund stark ausgeprägter Künstlichkeit der Erhebungssituation Rückschlüsse auf einen breiteren gesellschaftlichen Kontext nur eingeschränkt zulässig seien. Womöglich findet sich in der Literatur auch deshalb eine Vielzahl unterschiedlicher Abwandlungen und Erweiterungen des experimentellen Ansatzes, sodass man den Einzug des Experiments in die Soziologie auch als eine Abfolge verschiedener Kombinationen von Experiment und weiteren Datenquellen beschreiben könnte. Zu denken ist beispielsweise an die weit verbreiteten Feldexperimente oder auch die sich zunehmender Beliebtheit erfreuenden faktoriellen Survey-Experimente.

Ansinnen der vorliegenden Dissertation ist es, an diese Entwicklungen anzuknüpfen, den Fokus dabei allerdings auf bisher oftmals vernachlässigte methodische Aspekte zu richten. Konkret werden an verschiedenen empirischen Anwendungsbeispielen Möglichkeiten aufgezeigt, wie der Einbezug von (räumlichen) Kontextinformationen dazu genutzt werden kann, Effektheterogenität im sozialwissenschaftlichen Experiment zu berücksichtigen und so zu einer höheren externen Validität der Befunde beizutragen. Im ersten Kapitel werden die zu untersuchenden Überlegungen zu methodischen Problemen sozialwissenschaftlicher Experimente vor dem Hintergrund des Potential Outcomes Models vorgetragen und am Beispiel interdisziplinärer Erkenntnisse zum Umwelt- und Klimaschutz illustriert. Dabei wird deutlich, dass neben der Debatte um die interne und externe Validität experimenteller Studien insbesondere der adäquate Umgang mit Effektheterogenität von zentraler Bedeutung sowohl für die Analyse ursächlicher Mechanismen als auch für die Einordnung der Belastbarkeit beobachteter Effekte ist.

Im zweiten Kapitel werden diese Überlegungen anhand der Frage, inwiefern sich

Einstellungen in entsprechendem Verhalten niederschlagen, getestet. Am Beispiel der Befürwortung einer möglichen City-Maut in München wird untersucht, inwiefern das Umweltbewusstsein zu einer umweltgerechteren Befürwortung einer hypothetischen City-Maut beiträgt – auch wenn dadurch individuell mit finanziellen Mehrkosten zu rechnen ist. Dabei steht die explizite Modellierung theoretisch zu erwartender Heterogenität des experimentellen Stimulus über den Wertebereich eines Moderators hinweg im Fokus. Mithin wird der nicht-lineare Zusammenhang des Treatmenteffekts (die Höhe der zu erwartenden Mautgebühren) über moderierend wirkende Befragtenmerkmale (Einkommen und Umweltbewusstsein) modelliert und so ein erster Test der von Tutić et al. (2017) vorgeschlagenen Teststrategie zur Überprüfung der Low-Cost-Hypothese (im engeren Sinn eines variierenden Preiseffekts) vorgelegt.

In Kapitel 3 wird das Beispiel der Befürwortung einer City-Maut nochmals aus einer allgemeineren Perspektive aufgegriffen und demonstriert, wie der Einbezug des räumlichen Kontexts dazu genutzt werden kann, Heterogenität zu identifizieren und anschließend durch getrennte Analysen für verschiedene Subgruppen zu reduzieren. Bestehende Ansätze aus der Literatur werden durch die gekreuzte Betrachtung der relevanten Faktoren der Anwohnerschaft in einem von einer möglichen City-Maut betroffenen Gebiet (ja/nein) sowie die Frage des individuellen Autobesitzes (ja/nein) ergänzt. Diese Form der Heterogenität wäre durch gepoolte Analysen unbemerkt geblieben.

Im gemeinsam mit Katrin Auspurg und Andreas Schneck verfassten Forschungsbeitrag, der Kapitel 4 zugrunde liegt, wird ein gänzlich anderer empirischer Anwendungsfall untersucht. Dabei wird am Beispiel eines groß angelegten Feldexperiments zu ethnischer Diskriminierung auf dem deutschen Wohnungsmarkt erforscht, inwiefern typische Samplingverfahren zu einem verzerrten Abbild des zu untersuchenden Marktes führen und so womöglich die korrekte Identifikation des Ausmaßes von, aber auch der Risikofaktoren für Diskriminierung beeinträchtigen. Durch den Einbezug umfangreicher Kontextinformationen zur dem Feldexperiment zugrundeliegenden Online-Wohnungsplattform lässt sich zeigen, dass typische Samplingverfahren zwar durchaus zu einer Überrepräsentation von kleinen (privaten) Anbietern und solchen Angeboten, die schon seit länger Zeit inseriert wurden, führen. Allerdings wirken sich diese deskriptiven Verzerrungen nicht auf Zusammenhangsanalysen der Risikofaktoren für Diskriminierung aus. Welche Ausmaße zu erwartende Verzerrungen annehmen können wird anhand abschließend präsentierter Simulationen verschiedener Stichprobenziehungen aus der erhobenen Marktbeobachtung gezeigt. Vor diesem

Hintergrund lassen sich auch Befunde bisheriger Forschung einordnen und zum Teil womöglich kleinere Schwankungen der genauen Schätzungen unterschiedlicher Studien aufklären.

Insgesamt unterstreichen die drei empirischen Anwendungsbeispiele aus unterschiedlicher Perspektive die zentrale Bedeutung theoretischer Auseinandersetzungen mit der Kontextabhängigkeit von Treatmenteffekten. Die jeweils vorgeschlagenen Varianten des Einbezugs von Kontextfaktoren sollen dabei illustrieren, welche Möglichkeiten bestehen, verschiedenste Überlegungen zur Heterogenität von Treatmenteffekten im sozialwissenschaftlichen Experiment empirisch zu testen. In Anbetracht einer sich auf verschiedene (zumindest in Teilen auch sozialwissenschaftliche) Disziplinen erstreckenden Replikationskrise scheinen die vorgelegten Befunde auch darauf hinzudeuten, dass der äußere Kontext (in welchem Original- aber auch Replikationsstudie durchgeführt werden), Einfluss darauf haben kann, inwiefern tatsächlich mit den vermuteten Treatmenteffekten zu rechnen ist. Insofern erweist sich eine systematische Auseinandersetzung mit Effektheterogenität, so das Argument, auch von zentraler Bedeutung für die weitere Einordnung des theoretischen Erkenntnisfortschritts.

Inhaltsverzeichnis

Abbildungsverzeichnis **viii**

Tabellenverzeichnis **x**

1. Rahmenkapitel **1**

1.1. Einleitung 3

1.2. Sozialwissenschaftliche Experimente 7

1.2.1. Zentrale Bedeutung für den Test kausaler Hypothesen 7

1.2.2. Das Potential Outcomes Model 9

1.2.3. Methodische Debatte um die „Künstlichkeit“ von Experimenten und daraus resultierende Probleme der Validität von Befunden 21

1.2.4. Effektheterogenität als zentrale Herausforderung 24

1.3. Der Beitrag einer (stärkeren) Kombination mit Kontextinformationen 26

1.3.1. Zusammenfassung und Einordnung der eigenen Beiträge . . . 30

1.3.2. Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut 32

1.3.3. Support for city road tolls: a question of self-interest? 34

1.3.4. Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination 37

1.4. Fazit, Limitationen und Ausblick 40

Literaturverzeichnis 45

2. Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut **59**

2.1. Einleitung 61

2.1.1. Bisherige Forschung zur LCH 61

2.1.2. Der Beitrag eines neuerlichen Tests 64

2.2. Theorie . . . 65
2.2.1. Die mikroökonomische Modellierung der LCH . . . 65
2.2.2. Bisherige Prüfung der neuen Modellierung . . . 68
2.2.3. Das Anwendungsbeispiel einer City-Maut . . . 69
2.3. Daten und Methoden . . . 70
2.3.1. Operationalisierung . . . 70
2.3.2. Analysemodell . . . 74
2.4. Ergebnisse . . . 74
2.4.1. Verbesserte Teststrategie zur Prüfung der LCH . . . 75
2.4.2. Robustheitsanalysen . . . 76
2.4.3. Prüfung der LCH anhand der vormals üblichen Teststrategie . 80
2.5. Zusammenfassung . . . 80
2.6. Diskussion . . . 82
Literaturverzeichnis . . . 86
2.A. Anhang . . . 91

3. Support for city road tolls: a question of self-interest? 93
3.1. Urban road pricing . . . 95
3.2. State of research . . . 97
3.3. Theoretical arguments . . . 98
3.3.1. Beliefs on policy consequences . . . 99
3.3.2. Self-interest perspective . . . 100
3.4. Data and methods . . . 101
3.5. Results . . . 104
3.5.1. Descriptive results . . . 104
3.5.2. Institutional design . . . 107
3.5.3. Subgroup differences in preferred institutional design . . . 109
3.6. Summary . . . 112
3.7. Discussion . . . 113
References . . . 115
3.A. Appendix . . . 123

4. Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination 129
4.1. Introduction . . . 132
4.2. State of the Art: Sampling Strategies Used in Field Experiments . . . 134

4.3. Theoretical Background: Do Sampling Techniques Limit External Validity? . . . 136
4.3.1. Do Sampling Techniques Lead to a Biased Sample of Housing Units? . . . 138
4.3.2. Does Discrimination Vary with the Size of Suppliers or the Length of Advertisements? . . . 139
4.4. Data . . . 141
4.4.1. Sample of Housing Units . . . 141
4.4.2. Experimental Design . . . 142
4.4.3. Market Data . . . 143
4.4.4. Identification Strategy . . . 145
4.5. Results . . . 148
4.5.1. Is there a Sampling Bias? . . . 148
4.5.2. Does Sampling Bias Affect the Level of Discrimination? . . . 150
4.5.3. Does Sampling Bias Affect the Effects of Other Treatment Variables? . . . 155
4.5.4. Robustness Checks . . . 156
4.6. Summary . . . 158
4.7. Conclusions . . . 159
References . . . 163
4.A. Appendix . . . 171
4.A.1. Monte Carlo Simulations . . . 171
4.A.2. Regression Results . . . 173

Abbildungsverzeichnis

1.1. Schematische Darstellung der theoretischen Treatmenteffekte individueller Proband:innen einer hypothetischen Population mit und ohne Berücksichtigung eines Moderators 28
1.2. Schematische Darstellung der theoretischen Treatmenteffekte individueller Proband:innen einer hypothetischen Population mit und ohne Berücksichtigung verschiedener Subgruppen 29

2.1. Schematische Darstellung der anhand der Low-Cost-Hypothese zu erwartenden Zusammenhänge; dargestellt ist das vorhergesagte logarithmierte Ausmaß einstellungskonformen Verhaltens über die Einstellung (α; in Skalenpunkten; links), den Preis (p_a; in €; mittig) und das Einkommen (m; in €; rechts) 67
2.2. Beispielvignette mit Antwortskala (variierte Dimensionen unterstrichen) 73
2.3. Verteilung der Mautbewertungen (Darstellung basiert auf 4.317 Vignettenurteilen von 1.102 Personen) 75
2.4. Vorhergesagte Werte der logarithmierten Befürwortung einer Maut mit 95%-Konfidenzintervallen, dargestellt über das Umweltbewusstsein (links; wobei $m = 2000$, $p_a = 5$), die Mautgebühr (mittig; wobei $\alpha = 3$, $m = 2000$) und das Einkommen (rechts; wobei $\alpha = 3$, $p_a = 5$) . 76

3.1. Sample vignette with 11-point response scale 103
3.2. Toll scheme evaluation distribution 105
3.3. Mean evaluation with 95% confidence intervals by sub-group 106
3.4. Random-intercept model of toll scheme evaluation 108
3.5. Random-intercept models of toll scheme evaluations by sub-group . . 110

4.1. Lexis Diagram: Association Between the Advertisement Time and Likelihood of Being Sampled 139

4.2. Sample Inquiry (Translated Version, Experimentally Varied Dimensions Are Underlined) 144
4.3. Kernel Density Estimate of the Size of Supplier (Left Panel) and Advertisement Time (Right Panel) by Data Source (Sample vs. Market Data) 150
4.4. Local Polynomial Smooth Curves of the Discrimination Rates Against the Turkish (T) and German (G) Applicant Over the Size of Supplier (Left Panel) and Advertisement Time (Right Panel) 153
4.5. Multinomial Logistic Regressions, AMEs of Predictors with 95% Confidence Intervals for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment') 154
4.6. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time 157
4.A1. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time 179

Tabellenverzeichnis

1.1. Übersicht der eigenen Beiträge 31

2.1. Lineares Regressionsmodell zur Prüfung der Low-Cost-Hypothese . . 77
2.2. Robustheitsanalysen zur Absicherung der Befunde zur Low-Cost-Hypothese anhand verschiedener Regressionsmodelle der logarithmierten Befürwortung einer City-Maut 78
2.3. Lineare Regressionsmodelle der logarithmierten Befürwortung einer City-Maut zur Prüfung der Low-Cost-Hypothese anhand der vormals üblichen Teststrategie 81
2.A1. Übersicht der im Rahmen der Low-Cost-Hypothese vermuteten Zusammenhänge anhand partieller Ableitungen der Nachfragefunktion (vgl. Tutić et al., 2017, S. 657) 91
2.A2. Dimensionen und Levels der fiktiven Mautmodelle 92

3.A1. Overview of dimensions and levels experimentally varied in vignette descriptions 123
3.A2. Descriptive statistics 124
3.A3. Random-intercept models of support for city tolls, estimated for the entire sample and separately by sub-group 125
3.A4. Random-intercept models of support for city tolls, estimated separately for sub-groups 128

4.1. Sampling Techniques in E-Mail Correspondence Tests Published in 2010 – 2019 137
4.2. Descriptive Statistics on the Sample and Market Data 149
4.3. Response Patterns and Resulting Discrimination Rates 151
4.A1. Monte Carlo Simulations for Different Sampling Strategies 172

4.A2. Multinomial Logistic Regression Models of Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant'; Logit Coefficients and p-Values (in Parentheses) . . . 174
4.A3. Multinomial Logistic Regressions, AMEs of Predictors and p-Values (in Parentheses) for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment') as Reported in Figure 4.5 (in the Main Text) . . . 177
4.A4. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with p-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as Reported in Figure 4.6 (in the Main Text) . . . 178
4.A5. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with p-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as reported in Figure 4.A1 (in the Appendix) . . . 180

1. Rahmenkapitel

1.1. Einleitung

Der sechste Sachstandsbericht des Intergovernmental Panel on Climate Change verdeutlicht einmal mehr die großen Herausforderungen, die sich in der Bewältigung des Klimawandels stellen (IPCC, 2022). Demnach vermögen es bisherige Bestrebungen zur Eindämmung von Treibhausgasemissionen nicht, langfristige Erhöhungen der globalen Durchschnittstemperaturen entsprechend des Pariser Klimaabkommens von 2015 zu begrenzen. Das Ziel eines Anstiegs bis zum Jahr 2100 um nur 1,5 °C (bzw. maximal 2 °C) im Vergleich zum vorindustriellen Zeitraum (1850-1900) scheine demnach kaum mehr realistisch. Um diese Größenordnung doch noch erreichen zu können, benötige es einer Vielzahl tiefgreifender Maßnahmen, die neben der verstärkten Nutzung technologischer Innovationen insbesondere auch weitreichende Anpassungen hin zu klimaneutral(er)en Verhaltensweisen vorsehen (IPCC, 2022). Dass die hierzu nötigen Veränderungen der Rahmenbedingungen wiederum weitreichende gesellschaftliche Auswirkungen haben werden, liegt auf der Hand. Zu denken ist dabei etwa an das Auseinanderfallen von Verursachung und Betroffenheit von ökologischen Problemen und daraus resultierende Gerechtigkeitsfragen in der Verteilung von (Umwelt-)Belastungen, aber auch der nötigen Aufwendungen zu deren Beseitigung (für einen Überblick s. bspw. Preisendörfer, 2014). Weitere Aspekte umfassen etwa die durch klimatische Veränderungen regional verschärfte Knappheit lebensnotwendiger Ressourcen wie bspw. dem Zugang zu (sauberem Trink-)Wasser, sowie aus existenzbedrohender Knappheit erwachsende Verteilungskonflikte und infolge derer auch zunehmende Flucht- und Migrationsbewegungen (s. bspw. World Bank, 2016). Die Liste eindrücklicher Beispiele ließe sich beliebig fortführen. Nicht umsonst gelten der Klimawandel und damit zusammenhängende Probleme gemeinhin als *die* zentralen Herausforderungen des 21. Jahrhunderts (IPCC, 2022).

Zweifelsohne geht mit der weiteren Verschärfung des Klimawandels auch eine Vielzahl gesellschaftlicher Herausforderungen einher. In diesem Zusammenhang scheint der erst kürzlich erschienene gemeinsame Schwerpunkt der beiden Fachzeitschriften Nature Climate Change und Nature Human Behaviour (Antusch und Yan, 2022) besonders erwähnenswert. So liefert ein solches Sonderheft einen breiten Überblick über relevante Befunde und noch offene Forschungslücken. Genauso aufschlussreich erscheint darüber hinaus jedoch der Umstand, dass es sich um einen *gemeinsamen* Fokus der beiden Zeitschriften handelt – unterstreicht das doch auch strukturell die Bedeutung, die menschlichem Handeln nicht nur in der Verursachung, sondern

auch in der Bewältigung der Klimakrise beigemessen wird. So diskutiert etwa auch der IPCC-Bericht erstmals explizit die Bedeutung gesellschaftlicher und kultureller Dynamiken für die Klimafolgenanpassung, was perspektivisch auch neue Potentiale bietet für eine stärker verhaltensorientierte Klimafolgenforschung (Antusch, 2022; IPCC, 2022; Yan, 2022). An dieser Stelle wird eine Soziologie wichtig, die es einerseits vermag, akkurate Deskriptionen des Status quo zu liefern und andererseits, die ursächlichen Mechanismen, die den beobachteten Phänomenen zugrunde liegen, zu erforschen (für einen aktuellen Überblick über die Bandbreite empirisch-analytischer Soziologie siehe etwa die beiden Handbücher: Gërxhani et al., 2022; Manzo, 2021; für eine Übersicht wissenschaftlicher Gütekriterien einer so verstandenen Soziologie siehe etwa den jüngst erschienenen Beitrag von Otte et al., 2023; für eine Diskussion spezifisch soziologischer Beiträge zur Klimafolgenforschung sei insbesondere auf den Beitrag von Wiertz und de Graaf, 2022 verwiesen).

Gerade für die Analyse kausaler Fragestellungen eignen sich insbesondere experimentelle Ansätze (s. bspw. Berger und Wolbring, 2015; Gërxhani und Miller, 2022; Jackson und Cox, 2013; Veltri, 2021). Dabei werden als ursächlich angenommene Faktoren durch Forschende gezielt manipuliert, wodurch sichergestellt ist, dass der experimentelle Stimulus bzw. das Treatment einer zu beobachtenden Reaktion der Proband:innen zeitlich vorausgeht. So ist die Richtung des vermuteten Zusammenhangs identifizierbar. Durch die randomisierte Zuweisung von Stimuli auf Proband:innen (etwa über Versuchs- und Kontrollgruppen) lassen sich zudem Einflüsse eventueller Drittvariablen und damit alternativer Erklärungsmechanismen ausschließen. Gerade Laborexperimente erlauben gemeinhin ein besonders hohes Maß an Kontrolle, weshalb sie vielen Forschenden als „Goldstandard“ wissenschaftlicher Erkenntnis gelten (bspw. Falk und Heckman, 2009; Rubin, 2008).

Obwohl die Einsicht in die Notwendigkeit experimenteller Studien keineswegs neu ist, wie etwa die bereits 1928 in den Kölner Vierteljahresheften für Soziologie erschienenen Ausführungen Pitirim A. Sorokins zeigen,[1] verblieben Experimente in der Soziologie lange Zeit eher randständig (bspw. Gërxhani und Miller, 2022; Opp, 1969). Kritik richtete sich dabei häufig gegen eine als gering eingeschätzte Reali-

1 Im Wortlaut fordert er: „Ist Soziologie eine nomographische Wissenschaft, und erhebt sie den Anspruch, Regeln aufzustellen, die in exakter Weise die funktionalen und kausalen Zusammenhänge auf dem Gebiete der sozialen Erscheinungen beschreiben, so muß früher oder später der Zeitpunkt kommen, an dem ihre Probleme experimentell untersucht werden“ (Sorokin, 1928, S. 186). Die inzwischen bald einhundert Jahre zurück liegende Forderung hat auch heute nicht an Aktualität verloren (vgl. Berger und Wolbring, 2015).

tätsnähe – vor allem von Laborexperimenten. Im Labor beobachtetes Verhalten könne aufgrund der Künstlichkeit der Situation nur eingeschränkt auf Verhalten außerhalb des Labors übertragen werden. Die Verallgemeinerbarkeit der Befunde sei daher, trotz der hohen Kontrolle der Erhebungssituation, nur eingeschränkt möglich (für eine kritische Diskussion dieses Einwands siehe etwa Falk und Heckman, 2009; Findley et al., 2021; Thye, 2014).

Erst in den vergangenen Jahrzehnten kommen in der Soziologie verstärkt auch Experimente zum Einsatz. Dabei finden sich neben Laborexperimenten auch zunehmend interessante Kombinationen des experimentellen Ansatzes mit anderen Erhebungsmethoden (Gërxhani und Miller, 2022). Zu denken ist etwa an Feldexperimente, bei denen die randomisierte Setzung eines Stimulus nicht im Labor, sondern in natürlicher Umgebung erfolgt und die insofern eine vielversprechende Verknüpfung aus Experiment und Beobachtungsstudie darstellen (Wolbring und Keuschnigg, 2015). Eine weitere zunehmend verbreitete Variante besteht in der Kombination von Experiment und Befragungsstudie im faktoriellen Survey-Experiment. Dabei werden im Rahmen einer Befragung kurze systematisch variierte Beschreibungen (Vignetten) in randomisierter Zuordnung zur Bewertung vorgelegt.[2] Das erlaubt den Einbezug breiterer Bevölkerungsschichten, die für Laborexperimente oft nicht erreichbar wären (Auspurg und Hinz, 2015a,b; Mutz, 2011; Treischl und Wolbring, 2021).

Das Anliegen der vorliegenden Dissertation ist es, an diese Entwicklungen anzuknüpfen, den Fokus allerdings auf bisher wenig beachtete Möglichkeiten der Kombination von Experiment und weiteren Datenquellen zu legen. Dabei wird die These vertreten, dass das Potential solcher Kombinationen bei weitem noch nicht ausgeschöpft ist. Gerade der stärkere Einbezug von Kontextinformationen etwa stellt einerseits an sich bereits eine aufschlussreiche Erweiterung dar, findet soziale Interaktion und damit ein Kern soziologischer Forschung doch immer in Kontexten statt, die auch von der räumlichen Umgebung geprägt sind (für einen aktuellen Überblick etwa Small und Adler, 2019).[3] Andererseits kann der systematische Einbezug

2 Ähnlich dazu, in der Soziologie allerdings weit weniger verbreitet, ist das Choice Experiment. Während beim faktoriellen Survey-Experiment einzelne Vignetten zur Bewertung vorgelegt werden, sind beim Choice Experiment verschiedene solcher Optionen direkt gegeneinander abzuwägen (für eine Übersicht, siehe etwa Auspurg und Liebe, 2011; Liebe und Meyerhoff, 2021).

3 Die Frage nach dem Einfluss des räumlichen Kontexts auf soziale Bindungen hat in der Soziologie eine lange zurückreichende Tradition, findet sie sich doch etwa bereits bei Simmel (2016) und Blau (1977). Daran anknüpfend ist beispielsweise die Befassung mit der Entstehung und Struktur sozialer Netzwerke und deren Auswirkungen auf Tauschprofite zu nennen (Braun, 1993; Coleman, 1972). Aber auch in angrenzenden Disziplinen finden sich einfluss-

des räumlichen Kontexts auch gezielt dazu genutzt werden, oftmals vernachlässigten methodischen Problemen bestehender Ansätze zu begegnen. Konkret zielen die Bestrebungen darauf ab, an verschiedenen empirischen Anwendungsbeispielen praktische Möglichkeiten aufzuzeigen, wie durch den Einbezug des räumlichen Kontexts unterschiedliche Formen von Effektheterogenität in sozialwissenschaftlichen Experimenten berücksichtigt werden können, was letztendlich zu einer höheren externen Validität der Befunde beitragen soll.

Dabei sind die empirischen Anwendungsbeispiele bewusst so gewählt, dass sie möglichst eine Bandbreite unterschiedlicher Fragestellungen abdecken. Ihnen ist gemein, dass jeweils der Einbezug weiterer Kontextinformationen dazu genutzt wird, im Experiment auftretende Heterogenität des interessierenden Treatmenteffekts in geeigneter Weise zu berücksichtigen. Im Einzelnen umfasst das 1) die explizite Modellierung theoretisch erwarteter Heterogenität des Stimulus über den Wertebereich eines Moderators hinweg (in Form nicht-linearer Zusammenhänge; Kapitel 2), 2) die Verminderung von Heterogenität durch die getrennte Schätzung von Treatmenteffekten in Abhängigkeit moderierend wirkender Kontexteinflüsse (getrennte Analysen für Subgruppen; Kapitel 3) und 3) die Identifikation von Heterogenität in der Zusammensetzung des Samples (Kontrolle von Feldbedingungen; Kapitel 4).

Bevor ausführlicher auf den Beitrag der einzelnen Anwendungsstudien eingegangen wird, ist im Folgenden zunächst eine kurze Einführung in die Grundlagen des sozialwissenschaftlichen Experiments und seine Fundierung im Potential Outcomes Model zu geben (Abschnitt 1.2). Vor dem Hintergrund der Frage, als wie belastbar Ergebnisse experimenteller Designs anzusehen sind, wird auf die intensiv geführte Debatte um die Künstlichkeit von Experimenten und die daraus erwachsenden Implikationen für die interne und externe Validität von Befunden eingegangen, wobei schließlich der adäquate Umgang mit Heterogenität als eine der zentralen Herausforderungen für die Schätzung belastbarer Effekte ausgemacht wird. Daran anschließend wird dargelegt, welchen neuen Beitrag die Kombination bereits bekannter Varianten des Experiments mit Kontextinformationen leisten kann (Abschnitt 1.3). Dazu werden die empirischen Anwendungsbeispiele insbesondere in Bezug auf die sie verbindende methodische Fragestellung hin jeweils kurz zusammengefasst. Abschließend findet sich eine Diskussion der aufgezeigten Potentiale, aber auch der

reiche Arbeiten zum Wechselspiel aus räumlichem Kontext und individueller Handlungswahl, wie etwa an Schellings (1971) Befassung mit Segregationsmodellen deutlich wird. Für weitere Anwendungsbeispiele sei auf den Überblick in Small und Adler (2019) verwiesen.

Limitationen des gewählten Vorgehens sowie Desiderate für die künftige Forschung (Abschnitt 1.4).

1.2. Sozialwissenschaftliche Experimente

Die ideale Auswahl geeigneter Methoden sowie ihrer sinnvollen Kombination bestimmt sich zwar nach der zu untersuchenden Fragestellung und mag daher im konkreten Einzelfall unterschiedlich ausfallen. Dennoch ist festzuhalten, dass sich experimentelle Designs in besonderer Weise für den Test kausaler Hypothesen eignen (bspw. Jackson und Cox, 2013). Dabei werden als ursächlich angenommene Faktoren durch Forschende gezielt manipuliert. Wie eingangs bereits angerissen wurde, wird so sichergestellt, dass der experimentelle Stimulus der Reaktion von Proband:innen zeitlich vorausgeht, wodurch die Richtung des vermuteten Zusammenhangs identifizierbar ist. Durch die randomisierte Zuweisung von Stimuli auf Proband:innen (etwa über Versuchs- und Kontrollgruppen) lassen sich zudem Einflüsse eventueller Drittvariablen und damit alternativer Mechanismen ausschließen.

In der Praxis finden verschiedene Formen experimenteller Designs Anwendung, die diese charakteristischen Eigenschaften jeweils in unterschiedlicher Weise erfüllen (bspw. Berger und Wolbring, 2015). Im Folgenden wird ein knapper Überblick über die ihnen gemeinsamen formalen Grundlagen gegeben, bei dem an verschiedenen empirischen Beispielen aus der Literatur auf die methodischen Herausforderungen bei der Schätzung kausaler Effekte eingegangen wird. Es wird deutlich, dass Probleme der externen Validität von Experimenten oft mit Blick auf die Künstlichkeit der Situation verhandelt werden. Aus möglicher Heterogenität von Effekten erwachsende Einschränkungen der (externen) Validität erfahren hingegen in der Debatte um die Eignung verschiedener experimenteller Designs weit weniger Beachtung. Dabei erscheint es durchaus angeraten, die zum Teil implizite Annahme homogener Effekte empirisch zu prüfen. Heterogenität lässt sich, so die hier vertretene These, oft designbasiert bereits in der Konzeption des Experiments, etwa durch den gezielten und theoriegeleiteten Einbezug des räumlichen Kontexts berücksichtigen.

1.2.1. Zentrale Bedeutung für den Test kausaler Hypothesen

Das „klassische“ oder „echte“ Experiment zeichnet sich durch drei zentrale Merkmale aus (vgl. Berger und Wolbring, 2015, S. 40f., Diekmann, 2011, S. 337): i) die

kontrollierte Manipulation eines Stimulus oder auch Treatments, ii) die Bildung von (mindestens zwei) Gruppen, die sich bezüglich des Treatments unterscheiden, sowie iii) die randomisierte Zuteilung auf Gruppen.

Unter Manipulation versteht man dabei, dass das zu untersuchende Treatment der Kontrolle der Forschenden unterliegt. Die unabhängige Variable wird systematisch variiert, wodurch sich *ceteris paribus* eine eindeutige zeitliche Abfolge von vermuteter Ursache und beobachteter Reaktion ergibt. Somit kann auch die Richtung des Zusammenhangs bestimmt werden, zumindest unter der Annahme, dass Proband:innen das experimentell zugewiesene Treatment nicht bereits vorab antizipieren und ihre Reaktionen entsprechend anpassen.

Es werden mindestens zwei Gruppen gebildet: eine Versuchsgruppe, die das Treatment erhält, und eine Kontrollgruppe, die das Treatment nicht erhält. Mögliche Verzerrungsquellen, wie etwa im Hintergrund ablaufende allgemeine Trends oder einflussreiche Ereignisse während des Experiments sollten beide Gruppen gleichermaßen betreffen. Der Vergleich von Versuchs- und Kontrollgruppe erlaubt es sodann, die Wirkung des Treatments von der Wirkung veränderter Rahmenbedingungen zu unterscheiden und so eventuelle Alternativerklärungen auszuschließen. Der Effekt des eigentlichen Treatments zeigt sich mithin in der Differenz der beiden Gruppen.

Die für den experimentellen Ansatz sicherlich wichtigste Eigenschaft liegt jedoch in der randomisierten Zuweisung auf Versuchs- und Kontrollgruppe. Sie ist deshalb so zentral, weil durch die zufällige Zuweisung Unterschiede in der Zusammensetzung der beiden Gruppen als mögliche Alternativerklärung für beobachtete Reaktionen ausgeschlossen werden können. Die Idee dahinter ist, dass sich aufgrund der zufälligen Zuweisung eventuelle Unterschiede ausmitteln und so keine systematischen Unterschiede in den Merkmalen von Proband:innen der Versuchs- und der Kontrollgruppe verbleiben. Das gilt umso eher, je größer die Gruppengröße und desto geringer die Anzahl relevanter Drittvariablen (Berger und Wolbring, 2015, S. 41, Davies et al., 2008). Besonders hervorzuheben ist dabei, dass dies sowohl auf beobachtete als auch auf unbeobachtete Variablen zutrifft. Gerade diese Möglichkeit der designbasierten Kontrolle auch unbeobachteter (oder aufgrund praktischer Beschränkungen nicht beobachtbarer) Drittvariablen zeichnet das Experiment gegenüber anderen Methoden aus (Campbell und Stanley, 1966; Jackson und Cox, 2013).

1.2.2. Das Potential Outcomes Model

Die den angerissenen Überlegungen zum sozialwissenschaftlichen Experiment zugrunde liegende Kernidee lässt sich formal mithilfe des Potential Outcomes bzw. Counterfactual Models fassen. Es geht auf frühe Arbeiten von Neyman (1990) zum experimentellen Design zurück, die in Folge verschiedentlich aufgegriffen wurden. Besonders hervorzuheben sind dabei u.a. die Bemühungen zur Formalisierung des Modells durch Rubin (1974). Für einen umfassenden Überblick aus sozialwissenschaftlicher Perspektive sei etwa auf Morgan und Winship (2015) verwiesen, von denen auch der Kern der hier verwendeten Notation entlehnt ist. In der Literatur zu kausaler Inferenz finden sich vielfache Erläuterungen der grundlegenden Aspekte (so etwa jüngst bspw. Breen, 2022; Veltri, 2021). Orientiert an diesen Arbeiten soll hier die wesentliche Intuition des Ansatzes in Kürze wiedergegeben werden, da diese die Grundlage der weiteren Argumentation darstellt.

Schätzung von Average Treatment Effects

Bezeichnet man also den im Experiment zu untersuchenden Stimulus bzw. das Treatment mit D, so wird die Vermutung geprüft, dass D einen Effekt auf das Outcome Y hat. Namensgebender Kern des Potential Outcomes Models ist die Annahme, dass für jede Person (oder allgemeiner: jede Untersuchungseinheit) i ein potentielles Outcome für jede mögliche Ausprägung von D existiert ($Y^{D=d}$). Der individuelle Effekt von D auf Y ergibt sich für die $i$te Person als

$$\delta_i = y_i^d - y_i^{d'}. \tag{1.1}$$

Der individuelle Kausaleffekt bezeichnet also die Differenz der potentiellen Outcomes zweier Zustände des Treatments D, das hier in allgemeiner Form für verschiedene Werte d und d' dargestellt ist. Es lassen sich also ohne Weiteres auch komplexere Experimentaldesigns untersuchen, in denen mehrere Vergleiche zwischen unterschiedlichen Zuständen des Treatments D angestrebt werden. Für eine schlankere Notation genügt es allerdings vollkommen, für den Moment von einem binären Treatment auszugehen, das die möglichen Werte 1 und 0 annehmen kann. Die resultierenden potentiellen Outcomes wären in diesem Fall mit Y^1 und Y^0 gegeben. An dieser Stelle wird das „fundamental problem of causal inference“ (Holland, 1986, S. 947) offenkundig. So ist es schlicht nicht möglich, den individuellen Kausaleffekt direkt

zu beobachten, da eine Person i in der Realität nur jeweils einem Wert von D zugeordnet werden kann. Mithin kann nur eines der potentiellen Outcomes (entweder y_i^1 oder y_i^0) tatsächlich beobachtet werden − das jeweils andere Outcome verbleibt zwangsläufig kontrafaktisch.

Diesem Umstand wird mit dem Average Treatment Effect ATE begegnet:

$$E(\delta) = E(Y^1 - Y^0) = E(Y^1) - E(Y^0). \tag{1.2}$$

Der ATE ergibt sich aus dem durchschnittlichen Kausaleffekt in der untersuchten Population. Betrachten wir diesen nur für diejenigen Personen, die das Treatment tatsächlich erhalten haben ($D = 1$), so spricht man vom ATT

$$E(\delta|D = 1) = E((Y^1 - Y^0)|D = 1) = E(Y^1|D = 1) - E(Y^0|D = 1), \tag{1.3}$$

also dem Average Treatment Effect for the Treated. Analog dazu ergibt sich für diejenigen Personen in der Kontrollbedingung ($D = 0$), die das Treatment also nicht erhalten haben, der Average Treatment Effect for the Untreated (ATU) mit

$$E(\delta|D = 0) = E((Y^1 - Y^0)|D = 0) = E(Y^1|D = 0) - E(Y^0|D = 0). \tag{1.4}$$

Definitionsgemäß entspricht das beobachtete Outcome Y für jedes i dem potentiellen Outcome für das Treatment, das i tatsächlich zugewiesen wurde, was sich kompakt darstellen lässt als:

$$Y = DY^1 + (1 - D)Y^0. \tag{1.5}$$

Für die weitere Berechnung ist es nun zentral, dass Beobachtungen beider potentieller Outcomes vorliegen. Y^1 für diejenigen is, die dem Treatment ausgesetzt wurden und Y^0 für diejenigen is, die dem Treatment nicht ausgesetzt wurden. Man könnte also versucht sein, den ATE aus der bloßen Differenz der beobachteten Average Outcomes zu berechnen:

$$E(\delta) = E(Y|D = 1) - E(Y|D = 0). \tag{1.6}$$

Dabei stellt sich die Frage, inwiefern dieser naive Schätzer des ATE (im Folgenden daher mit ATE* bezeichnet) ein geeignetes Maß des wahren ATE darstellt. Um das zu beurteilen, lässt sich der in Gleichung (1.2) gegebene ATE durch Dekomposition

zunächst schreiben als

$$\begin{aligned} E(\delta) =&[\pi E(Y^1|D=1) + (1-\pi)E(Y^1|D=0)]- \\ &[\pi E(Y^0|D=1) + (1-\pi)E(Y^0|D=0)]. \end{aligned} \tag{1.7}$$

Dabei gibt π, definiert als $E(D)$, den Anteil derjenigen is an, die das Treatment erhalten haben. $(1-\pi)$ bezeichnet entsprechend den Anteil is in der Kontrollgruppe, die das Treatment nicht erhalten haben. Der ATE setzt sich damit aus fünf Komponenten zusammen − im Einzelnen umfasst er den Anteil is, die das Treatment erhalten haben, sowie die vier potentiellen Outcomes.[4] Es braucht nun unweigerlich weitere Annahmen, um die unbekannten Größen zu schätzen und so den erwarteten Kausaleffekt des Treatments bestimmen zu können. Deutlich wird das insbesondere durch weitere Umformung des ATE*:

$$\begin{aligned} E(Y^1|D=1) - E(Y^0|D=0) =&E(\delta) + [E(Y^0|D=1) - E(Y^0|D=0)] \\ &+ (1-\pi)[E(\delta|D=1) - E(\delta|D=0)]. \end{aligned} \tag{1.8}$$

Der naive Schätzer ATE* konvergiert also zu einer einfachen Differenz, die tatsächlich dem wahren ATE, $E(\delta)$, plus zwei Fehlertermen entspricht. Im Einzelnen handelt es sich dabei um den sogenannten baseline bias und um den differential treatment effect bias.

Der baseline bias, $[E(Y^0|D=1) - E(Y^0|D=0)]$, ergibt sich dabei in Abwesenheit des Treatments aus der Differenz des potentiellen Outcomes zwischen denjenigen, die das Treatment tatsächlich erhalten haben und denjenigen, die es nicht erhalten haben. Welcher Unterschied im Outcome wäre also zwischen den is in Treatment- und Kontrollgruppe beobachtbar, hätten sie das Treatment nicht erhalten? Der differential treatment effect bias, $(1-\pi)[E(\delta|D=1) - E(\delta|D=0)]$, ergibt sich hingegen, wenn sich der Effekt für diejenigen, die das Treatment erhalten haben, vom Effekt unterscheidet, den diejenigen hätten, die das Treatment nicht erhalten haben, wenn sie es erhalten hätten (multipliziert mit dem Anteil derjenigen, die das Treatment nicht erhalten haben).

ATE* stellt einen geeigneten Schätzer für den ATE dar, wenn sowohl baseline

4 Mithilfe des Gesetzes der großen Zahlen kann gezeigt werden, dass das durchschnittliche beobachtete Outcome Y mit steigender Sample-Größe $N \to \infty$ zum wahren durchschnittlichen Outcome unter Treatment-Bedingungen für diejenigen in der Treatmentgruppe, $E(Y^1|D=1)$, sowie zum wahren durchschnittlichen Outcome unter Kontrollbedingungen für diejenigen in der Kontrollgruppe, $E(Y^0|D=0)$, konvergiert (Morgan und Winship, 2015, S. 57f.).

bias als auch differential treatment effect bias null sind. Und das ist genau dann der Fall, wenn keine systematischen Unterschiede zwischen den Merkmalsverteilungen in Treatment- und Kontrollbedingung auftreten, die sich nicht auf das Treatment selbst zurückführen lassen. Der bereits angesprochene zentrale Vorteil des experimentellen Ansatzes besteht darin, dass hier (abgesehen von kleineren Zufallsschwankungen) eben jene Gleichverteilung beliebiger Merkmale Z durch randomisierte Zuteilung auf Treatment- und Kontrollgruppe (zumindest bei entsprechender Gruppengröße und überschaubarer Anzahl möglicher Zs) näherungsweise sichergestellt wird. Je größer die Gruppen und je geringer die Anzahl möglicher Störgrößen, desto eher sollten Unterschiede zwischen den Gruppen verschwinden.

Illustration am Beispiel: Die Wirksamkeit „grüner Nudges“

Die Intuition, die der Schätzung eines Kausaleffekts im Experiment zugrunde liegt, lässt sich etwa am Beispiel „grüner Nudges“ illustrieren. Nudging gilt als eine vielversprechende Möglichkeit wünschenswertes Verhalten anzuregen (Thaler und Sunstein, 2021). Dabei werden bestimmte, in diesem Fall ökologisch nachhaltige, Verhaltensentscheidungen durch kleine „Anstupser“ (beispielsweise in bestimmter Weise dargestellte Informationen, passgenau bereitgestelltes Feedback, veränderte Default-Einstellungen, etc.) begünstigt, ohne jedoch andere (nicht wünschenswerte) Entscheidungen zu sanktionieren oder gar zu verbieten.[5]

Nehmen wir etwa an, dass die Wirksamkeit einer Informationskampagne zur Förderung ressourcensparenden Verhaltens untersucht werden soll. Es wird vermutet, dass allein schon passgenaue Informationen in Form individuellen Feedbacks über den eigenen Verbrauch (auch im Vergleich zum Verbrauch anderer) genügen, um zu Reduktionen des Ressourcenverbrauchs beizutragen. Um diese Vermutung zu überprüfen, bietet sich ein (feld-)experimentelles Design an, bei dem der tatsächliche Verbrauch gemessen wird. Das Treatment besteht sodann in der Information der Proband:innen sowohl über ihren jeweils eigenen Verbrauch als auch darüber, inwiefern sie mit diesem Verbrauch im Vergleich zu anderen über- oder unterdurchschnittlich abschneiden (wodurch letztlich eine deskriptive Norm beschrieben wird). Proband:innen in der Kontrollgruppe erhalten diese Information nicht. Die Zuteilung von Proband:innen auf Treatment- und Kontrollgruppe erfolgt randomisiert, sodass

5 Für eine kritische Diskussion, inwiefern Nudging dieses Versprechen eines „libertarian paternalism“ konzeptionell tatsächlich einzulösen vermag, ist beispielsweise auf den Beitrag von Hausman und Welch (2010) zu verweisen.

sich Proband:innen der beiden Experimentalgruppen im Schnitt lediglich darin unterscheiden, ob sie das Treatment erhalten haben oder nicht. Sinkt der Verbrauch in der Treatmentgruppe nachdem die Information über den Verbrauch erfolgte, so lässt sich dieser Effekt sehr wahrscheinlich auf das Treatment zurückführen.

Empirisch wurden Effekte solcher Informations-Treatments beispielsweise für den (Warm-)Wasserverbrauch beim Duschen (Tiefenbeck et al., 2018, 2019), den Energieverbrauch im Haushalt (Allcott, 2011, 2015) oder die Wiederverwendung von Handtüchern in Hotelzimmern (Bohner und Schlüter, 2014; Goldstein et al., 2008) untersucht. Während die initialen Ergebnisse zum Teil durchaus vielversprechend wirken, treten bei genauerem Hinsehen verschiedene Probleme zu Tage. So berichten etwa Goldstein et al. (2008) in einer vielbeachteten Studie von deutlich gesteigerten Wiederverwendungsraten von Handtüchern, wenn Hotelgäste anstatt eines allgemeinen Appells zum Schutze der Umwelt einen Hinweis (auf die deskriptive Norm) erhalten, dass ein Großteil der anderen Hotelgäste Handtücher auch wiederverwenden. Die Befunde der ursprünglich im Südwesten der USA durchgeführten Feldexperimente konnten Bohner und Schlüter (2014) allerdings für Deutschland trotz gleichen Prozederes nicht replizieren. Für die unterschiedlichen Befunde kommen verschiedene Erklärungen in Frage, unter anderem, dass die Replikationsstudie in einem anderen kulturellen Kontext durchgeführt wurde als das Original, worauf etwa die durchweg höheren Wiederverwendungsraten (auch in der Kontrollgruppe) hindeuten (Bohner und Schlüter, 2014).

Bei einer Untersuchung des Opower energy conservation programs kommt Allcott (2011, 2015) zu gemischten Ergebnissen. Im Rahmen dieses umfassenden Feldexperiments wurden bis Februar 2013 an 111 Standorten in den gesamten USA zusammen mit der Abrechnung sogenannte Home Energy Reports an 8,6 Millionen Haushalte bei 58 verschiedenen Energieversorgungsunternehmen verschickt.[6] Das Treatment umfasste Feedback zum eigenen Verbrauch, kombiniert mit Informationen über den Verbrauch ähnlicher Haushalte in der Wohnumgebung und führte im Schnitt zu Reduktionen des Stromverbrauchs von etwa 2%. Das entspricht in der Größenordnung ungefähr der Reduktion, wie man sie durch eine Preiserhöhung um 11-20% kurzfristig (oder rund 5% langfristig) erwarten würde. Dabei ist wichtig zu betonen, dass ein solcher vergleichsweise starker Treatmenteffekt nicht für alle Haushalte gleichermaßen zu beobachten ist. Gerade Haushalte mit zuvor hohem Verbrauch zeigen mit

6 Das frühere Papier ist beschränkt auf die ersten 17 bis Ende 2009 implementierten Opower-Programme (knapp 600.000 Haushalte) in verschiedenen Regionen der USA (Allcott, 2011).

Reduktionen von etwa 6,3% stärkere Reduktionen als Haushalte mit bereits zuvor niedrigerem Verbrauch, bei denen ein Rückgang von lediglich 0,3% zu beobachten ist (Allcott, 2011).

Darüber hinaus zeigen sich im weiteren Verlauf des Opower energy conservation programs abnehmende Treatmenteffekte. Anhand früher Erhebungen ließen sich zunächst sehr hohe Reduktionen des Energieverbrauchs prognostizieren, die jedoch mit weiterer Ausbreitung des Programms abnahmen. So ist etwa eine nur auf den ersten 10 Standorten beruhende Schätzung des durchschnittlichen Treatmenteffekts der Home Energy Reports für die später hinzugekommenen 101 Standorte um etwa 0,41 bis 0,66 Prozentpunkte überschätzt. Diese systematische Verzerrung ist etwa auf eine positive Selektion von zunächst insgesamt relativ umweltfreundlichen Regionen und Haushalten mit einem hohen Ausgangsniveau des Energieverbrauchs zurückzuführen (Allcott, 2015).

Die bereits angesprochenen Feldexperimente zu möglichen Energieeinsparungen beim Duschen deuten auf ein ähnliches Muster. Durch experimentelle Variation der Verbrauchsanzeige auf einem sogenannten Smart Meter wurde Proband:innen live Feedback zu Ihrem Wasser- und Energieverbrauch während des Duschens gegeben.[7] Das zugrundeliegende Sample umfasst 700 Ein- bis Zwei-Personen-Haushalte in der Schweiz, die bereits zuvor an einer Studie zu Smart Metering teilnahmen und der Datenübermittlung ihres Duschverhaltens an die Forschenden zustimmten. Tiefenbeck et al. (2018) finden hier eine Reduktion von rund 22% des Energieverbrauchs, wobei dieser Einsparungseffekt wiederum für diejenigen mit zuvor hohem Verbrauch stärker ausfällt.

In einer Folgestudie wurden solche Smart Meter in sechs Schweizer Hotels mit insgesamt 265 Hotelzimmern installiert und so Daten von (bereinigt) 19.596 Wasserentnahmen im Zeitraum zwischen Februar und April 2016 gesammelt (Tiefenbeck et al., 2019). So konnte mögliche (Selbst-)Selektion, wie sie bei der Untersuchung des Freiwilligensamples zu vermuten ist, vermieden werden und zugleich eine Situation beobachtet werden, in der zudem keine finanziellen Anreize zu sparsamem Verhalten existieren. Im direkten Vergleich zur vorigen Studie beträgt die mittlere Reduktion für diejenigen Hotelgäste mit Live-Feedback nurmehr 11,4% des benötigten Energieverbrauchs. Der geschätzte Treatmenteffekt hat sich mithin fast halbiert, ist jedoch

7 Ein Smart Meter ist ein kleines Gerät, das hier zwischen Armatur und Brause installiert wird und so den Wasser- und Energieverbrauch während des Duschvorgangs erfassen und live auf einem kleinen Display anzeigen kann (siehe Tiefenbeck et al., 2018, Abb. 1).

weiterhin in substanzieller Größe zu beobachten (Tiefenbeck et al., 2019). Dennoch bleibt die Frage, inwiefern diese Befunde persistent sind und sich auch auf andere Situationen übertragen lassen. Immerhin handelt es sich bei wenigen Übernachtungen in einem Hotel mit Smart Meter in der Dusche womöglich um eine spezielle Situation, die nicht ohne Weiteres Aufschluss darüber gibt, inwiefern derlei Feedback-Tools dauerhaft und auch in anderen Situationen (etwa beim Heizenergieverbrauch in der Wohnung), zu niedrigerem Energieverbrauch beizutragen vermögen.

Der Schätzung zugrunde liegende Annahmen

In der bisherigen Argumentation sind wir von homogenen, für alle Proband:innen gleichermaßen zu erwartenden Treatmenteffekten ausgegangen. Was aber, wenn diese Annahme überhaupt nicht plausibel zu treffen ist? Immerhin ließe sich durchaus vermuten, dass gerade umweltbewusstere Personen ceteris paribus eher daran interessiert sind, ihren Ressourcenverbrauch zu reduzieren und daher womöglich auch stärker auf ein solches Informations-Treatment reagieren. Umgekehrt dürften Personen mit ohnehin schon geringem Verbrauch (sei es aufgrund eines hohen Umweltbewusstseins oder geringer finanzieller Mittel) weniger Potential für Verbrauchsreduktionen aufweisen als Personen mit vormals hohem Verbrauch. So deuten diese Überlegungen und die aufgeführten Befunde sowohl konzeptionell als auch empirisch darauf hin, dass der Kontext, in dem ein (Feld-)Experiment durchgeführt wird, durchaus eine Rolle spielt. Allgemeiner gefasst kommt beispielsweise Breen (2022, S. 283) zu dem Schluss, dass wohl die meisten in den Sozialwissenschaften untersuchten Effekte heterogen sind, die Identifikation von Heterogenität allerdings trotz einiger Beiträge auch aus der Soziologie (Breen et al., 2015; Xie et al., 2012; Zhou und Xie, 2020) ein schwieriges Problem bleibe.

Welche Konsequenzen ergeben sich daraus nun für die valide Schätzung kausaler Zusammenhänge im sozialwissenschaftlichen Experiment? Für die Schätzung des ATEs ist es nötig, dass verschiedene Bedingungen erfüllt sind. Hierbei ist zunächst auf die sogenannte Stable Unit Treatment Value Assumption (kurz: SUTVA; Rubin, 1980, 1986) einzugehen. Sie besteht im Kern aus zwei Komponenten, die besagen, dass i) die potentiellen Outcomes für Proband:in i nicht vom Treatmentstatus anderer Proband:innen j oder dem Zuweisungsmechanismus des Treatments abhängen und ii), dass keine unterschiedlichen Versionen desselben Treatments vorliegen, die

zu unterschiedlichen potentiellen Outcomes führen.[8] Obwohl zentral für den vorliegenden Ansatz, gibt es eine Reihe von Situationen, in denen diese Annahme wohl nicht erfüllt ist (für eine Diskussion daraus resultierender Implikationen, siehe beispielsweise Morgan und Winship, 2015, S. 48-52, VanderWeele und Hernan, 2013; VanderWeele et al., 2014).

Ein klassisches Beispiel für die Verletzung der ersten Komponente von SUTVA, also der Annahme, dass keine Interferenz zwischen den Proband:innen auftritt, liegt etwa bei der Analyse der Wirksamkeit von Impfkampagnen vor (Hudgens und Halloran, 2008; Ross, 1916). Dabei hängt die Wahrscheinlichkeit, dass sich Person i mit einer ansteckenden Krankheit infiziert von dem Anteil geimpfter Personen j in der Bevölkerung ab. In Folge wird die Bereitschaft, sich selbst einer Impfung zu unterziehen mit steigendem Anteil anderer geimpfter Personen und damit sinkendem individuellen Risiko einer Ansteckung für Person i abnehmen. In einer solchen Situation sind die potentiellen Outcomes für i und j nicht unabhängig voneinander. Ähnliches mag auch bei anderen Kollektivgutproblematiken etwa zum Umwelt- und Klimaschutz auftreten, bei denen die potentiellen Outcomes der Entscheidungen einer Person i oftmals auch vom Verhalten anderer abhängen.

Untersucht man beispielsweise die Erfolgsaussichten eines operativen Eingriffs und geht davon aus, dass verschiedene Chirurg:innen den Eingriff durchführen und diese unterschiedlichen Versionen des Treatments zu unterschiedlichen potentiellen Outcomes führen, ist die zweite Komponente von SUTVA verletzt (VanderWeele und Hernan, 2013). Eine naheliegende Möglichkeit, damit umzugehen, besteht in der Neudefinition des Treatments und dabei die Version zu integrieren. Man würde also nunmehr die Wirkung eines umfangreicheren Sets verschiedener Treatmentlevel mit jeweils klar definierten potenziellen Outcomes untersuchen. Auf diese Weise kann die Annahme, dass keine verschiedenen Versionen des Treatments vorliegen, aufrechterhalten werden. Dabei ist allerdings zu beachten, dass eine solche Neudefinition des Treatments nicht in allen Situationen eine inhaltlich aufschlussreiche Alternative darstellt oder für manche Fragestellungen womöglich aufgrund begrenzter Datenverfügbarkeit nicht praktikabel erscheint. So macht es beispielsweise einen substanziellen Unterschied, ob man sich für die insgesamt zu erwartenden Erfolgsaussichten eines operativen Eingriffs interessiert oder für die Erfolgsaussichten des Eingriffs,

8 Die „no interference“-Annahme geht Rubin (1980, S. 591) zufolge auf Cox (1958, S. 19) zurück. Die Annahme, dass keine unterschiedlichen Versionen des Treatments vorliegen, schreibt er Neyman (1935) zu.

wenn er von bestimmten Chirurg:innen durchgeführt wird. Ein weiteres Problem besteht dann, wenn Informationen über die Version des Treatments nicht erfasst werden und das eigens neu definierte Treatment letztlich empirisch nicht beobachtet wird (VanderWeele und Hernan, 2013). Aufschluss über die Version des Treatments könnte sich etwa auch aus dem Kontext ergeben, in dem das Treatment gesetzt wird. So mag beispielsweise eine politische Maßnahme zur Verringerung mobilitätsbedingter Emissionen regional unterschiedlich starke Auswirkungen auf bisherige Mobilitätsgewohnheiten haben. Mithin ergeben sich für verschiedene Regionen jeweils unterschiedliche Versionen desselben Treatments, die sich durch den gezielten Einbezug des räumlichen Kontexts differenzieren ließen.

Strategien zum Umgang mit Heterogenität

Steht die Erforschung grundlegender, als universell angenommener, „Konstanten" menschlichen Handelns im Vordergrund, so kann wohl tatsächlich von homogenen Effekten ausgegangen werden. In einer solchen Situation dürfte es keine Rolle spielen, nach welchen Kriterien Proband:innen für die Analyse rekrutiert werden. Randomisierung auf Treatment- und Kontrollgruppe sollte genügen, um belastbare Ergebnisse sicherzustellen. Entsprechend ist in solchen Fällen auch davon auszugehen, dass sich Treatmenteffekte in unterschiedlichen Kontexten und mit unterschiedlichen Samples replizieren lassen (Kohler et al., 2019).

Wie bereits ausgeführt, ist in vielen anderen Anwendungsfällen in den Sozialwissenschaften hingegen mit Effektheterogenität zu rechnen. Werden dabei die aus vorliegender Heterogenität resultierenden Implikationen für die Schätzung kausaler Zusammenhänge vernachlässigt, laufen Forschende Gefahr, verzerrten Ergebnissen aufzusitzen (die sich womöglich später auch nicht replizieren lassen, wie etwa Gelman (2015) argumentiert). Für die Analyse von Fragestellungen, bei denen nicht plausibel von homogenen Effekten ausgegangen werden kann, schlagen beispielsweise Kohler et al. (2019, S. 161) drei mögliche Strategien zum weiteren Vorgehen vor. So könne i) die Untersuchung auf spezielle, in sich homogene, Populationen beschränkt werden (deren Untersuchung als solches jedoch einen aufschlussreichen Beitrag zum Erkenntnisfortschritt leisten sollte), ii) die Untersuchung gezielt auf die Modellierung von Interaktionen (und damit die Erforschung der Ursachen von Heterogenität) gerichtet werden oder iii) eine Kombination deskriptiver und kausalanalytischer Ansätze dazu genutzt werden, mithilfe der Verteilungen individueller Treatmenteffekte einen Populations-ATE einer wohldefinierten, endlichen Populati-

on zu beschreiben.

In diese Richtung weisen auch Forderungen nach der gezielt theoriegeleiteten Entwicklung von Forschungsdesigns zur Identifikation von Heterogenität (Bryan et al., 2021; Tipton, 2013a; Tipton und Peck, 2017; Tipton et al., 2019). Eine solche Fokussierung könnte letztlich dazu beitragen, ursächliche Mechanismen, die beobachteten Zusammenhängen zugrunde liegen, genauer zu verstehen und so zur Theorieentwicklung beitragen. Wenig belastbare Verallgemeinerungen, die auf Erkenntnissen beruhen, die in Wirklichkeit nur für spezielle Populationen gelten oder bestimmte Bedingungen voraussetzen, ließen sich durch eine tiefgreifendere Befassung mit relevanten Moderatoren vermeiden. Mithin sollte bereits a priori festgelegt werden, für welche Subgruppen unter welchen Umständen welche Treatmenteffekte zu erwarten sind und anschließend ein für diese Anforderungen geeignetes Design vorgeschlagen werden (beispielsweise Gelman, 2015; Tipton et al., 2019).

Dazu erscheint es sinnvoll, auch in der Notation deutlich zu machen, für wen welche Treatmenteffekte geschätzt werden (vgl. Tipton et al., 2019). So beschreibt der SATE

$$E_S(\delta) = E_S(Y^1 - Y^0) = E_S(Y^1) - E_S(Y^0) \tag{1.9}$$

den Sample Average Treatment Effect, wobei das Subskript S darauf hindeutet, dass hier der ATE über alle $i = 1, \ldots, n$ Proband:innen im Sample angegeben ist. Für homogene Treatmenteffekte sollte der aus verschiedenen Samples geschätzte SATE sich abgesehen von kleineren Zufallsschwankungen nicht unterscheiden und den ATE in der zugrundeliegenden Population im Schnitt unverzerrt repräsentieren. Liegen hingegen variierende Treatmenteffekte vor, ist davon auszugehen, dass sich der SATE für verschiedene Samples und Populationen unterscheidet (beispielsweise Imai et al., 2008).

Vielversprechend erscheint daher die zunehmende Integration von Experimenten in Zufallssamples, was die direkte Schätzung des ATE in der Inferenzpopulation, dem PATE, erlaubt (Mutz 2011). Der PATE

$$E_P(\delta) = E_P(Y^1 - Y^0) = E_P(Y^1) - E_P(Y^0) \tag{1.10}$$

beschreibt den ATE für alle $i = 1, \ldots, N$ in der Inferenzpopulation, wobei das Subskript P wiederum auf die Schätzung des ATE in der Inferenzpopulation verweist. Während die unverzerrte Schätzung des SATE qua Randomisierung im Experiment

sichergestellt werden kann, erfordert die Schätzung des PATE darüber hinaus die zufällige Ziehung des Samples aus der Inferenzpopulation. Dabei muss angenommen werden, dass Non-Response im realisierten Sample unabhängig ist von jedweder Ursache von Treatmentheterogenität. Basiert die Schätzung des PATE schlussendlich nicht auf einem Zufallssample aus der Inferenzpopulation, kann nicht ausgeschlossen werden, dass Ergebnisse durch Selektionsbias verzerrt sind (Allcott, 2015; Tipton, 2013a; Tipton et al., 2019). Erschwerend ist zu beachten, dass in der Praxis etwa aufgrund von Responseraten (deutlich) unter 100% keine echten Zufallssamples existieren, mithin nicht qua Design sichergestellt werden kann, dass gleiche beziehungsweise zumindest angebbare Auswahlwahrscheinlichkeiten zu einer zufälligen Zusammensetzung des Samples führen (Kohler et al., 2019). Dabei spielt es an sich keine Rolle, ob schon von vornherein kein zufälliges Sample gezogen wurde oder ein solches durch selektive Non-Response entstanden ist (wobei das Ausmaß der Selektivität im Einzelfall durchaus unterschiedlich ausfallen dürfte).

Eine Möglichkeit, Variationen des Treatmenteffekts zwischen verschiedenen Subgruppen zu untersuchen, besteht sodann in der Dekomposition des PATE in mehrere ATEs einzelner Subgruppen

$$E_P(\delta) = E_{C_1}(\delta)\pi_1 + E_{C_2}(\delta)\pi_2 + \cdots + E_{C_H}(\delta)\pi_H = \sum E_{C_h}(\delta)\pi_h, \tag{1.11}$$

wobei $E_{C_h}(\delta)$ den CATE, also den für die Subgruppe $h = 1, \ldots, H$ konditionalen ATE, angibt (Tipton et al., 2019). Der Populationsanteil in Subgruppe h ist mit π_h bezeichnet (wobei $\sum \pi_h = 1$). Anders ausgedrückt setzt sich der PATE aus der gewichteten Summe der CATEs zusammen.

Aufschlussreich ist sodann die genauere Betrachtung der CATEs verschiedener Subgruppen. Unterscheiden sich diese nicht voneinander, kann von Homogenität ausgegangen werden. Der geschätzte CATE jedes Subsamples entspricht in diesem Fall (abgesehen von kleineren Zufallsschwankungen) im Schnitt dem PATE, für dessen unverzerrte Schätzung mithin kein Zufallssample erforderlich ist.

Unterscheiden sich die CATEs zwischen den Subgruppen (was auf Heterogenität des Treatmenteffekts deutet), die anteilige Besetzung der verschiedenen Subgruppen entspricht jedoch denen in der Inferenzpopulation, liefert der so geschätzte SATE (wiederum von Zufallsschwankungen abgesehen) eine unverzerrte Schätzung des PATE.

Unterscheiden sich sowohl die CATEs zwischen den Subgruppen als auch die an-

teilige Besetzung der Subgruppen im Sample von den Anteilen in der Inferenzpopulation, liefert der SATE keine unverzerrte Schätzung für den PATE. Der SATE lässt sich in diesem Fall also nicht auf die Inferenzpopulation übertragen. Liegen jedoch genug Beobachtungen einer inhaltlich relevanten Subgruppe vor, könnte man am CATE als Schätzung des ATE dieser speziellen Subgruppe interessiert sein. Sofern angenommen werden kann, dass innerhalb der Subgruppe keine Heterogenität vorliegt, liefert ein solches Vorgehen von kleineren Zufallsschwankungen abgesehen im Schnitt eine unverzerrte Schätzung. Das zeigt etwa die weitere Dekomposition des CATE einer bestimmten Subgruppe h:

$$E_{C_h}(\delta) = [E_{CC_{h1}}(\delta)\pi_{h1} + E_{CC_{h2}}(\delta)\pi_{h2} + \cdots + E_{CC_{hK}}(\delta)\pi_{hK} = \sum E_{CC_{hk}}(\delta)\pi_{hk}]/\pi_h. \tag{1.12}$$

Der CATE kann dabei in $k = 1, \ldots, K$ Sub-Subgruppen ATEs zerlegt werden (so gesehen konditional-konditionale ATEs bzw. CCATEs; Tipton et al., 2019). Der Populationsanteil in diesen Sub-Subgruppen ist mit π_{hk} angegeben (wobei $\sum \pi_{hk} = \pi_h$). Analog zur Dekomposition des PATE in mehrere Subgruppen CATEs kann davon ausgegangen werden, dass Homogenität innerhalb der Subgruppe h vorliegt, wenn die Sub-Subgruppen CCATEs innerhalb von h sich nicht voneinander unterscheiden. Liegt hingegen Heterogenität innerhalb der Subgruppe vor, muss die anteilige Besetzung der Sub-Subgruppen im Sample der Subgruppe derjenigen der Subgruppe in der Inferenzpopulation entsprechen. Andernfalls liefert der CATE keine unverzerrte Schätzung für den ATE dieser Subgruppe. Das verdeutlicht einmal mehr den konzeptionellen Vorteil von Zufallssamples selbst bei der Untersuchung von Experimenten (Tipton et al., 2019). Durch die zufällige Zusammensetzung des realisierten Samples sollten systematische Verzerrungen im Schnitt ausgeschlossen sein.

Im Prinzip lässt sich das Vorgehen iterativer Dekomposition beliebig fortführen. Zu bedenken ist dabei jedoch, dass es mit Blick auf eine aufschlussreiche Interpretation wohl sinnvoll erscheint, sich auf inhaltlich bedeutsame Subgruppen zu beschränken, anstatt Effekte für sehr spezifisch zugeschnittene Kleinstgruppen zu berichten. Darüber hinaus ist selbst bei insgesamt großen Zufallssamples auch auf eine entsprechende Besetzung der Subgruppen zu achten, da sonst insbesondere für selten vorkommende Subgruppen nicht ausreichend statistische Power vorliegt. Hier mag eine Lösungsstrategie gezieltes Oversampling einer solchen Subgruppe sein. Dadurch ließe sich der CATE für diese Subgruppe belastbar schätzen. Eine unverzerrte

Schätzung des PATE ist dann allerdings nurmehr mithilfe von Poststratifizierung möglich, wofür es wiederum belastbarer Informationen zur Zusammensetzung der Inferenzpopulation bedarf (Tipton, 2013a,b; Tipton und Peck, 2017; Tipton et al., 2019).

Jenseits dessen betonen etwa Kohler et al. (2019), dass Überlegungen zur Heterogenität von Treatmenteffekten nicht automatisch zu Bemühungen zur Schätzung des PATE führen sollten. Letztlich hänge es im Einzelnen von der konkreten Forschungsfrage ab, ob ein durchschnittlicher Treatmenteffekt über die gesamte Inferenzpopulation oder vielmehr die gezielte Schätzung von Treatmenteffekten für verschiedene Subgruppen aufschlussreich erscheint. Dabei zielten Analysen nicht nur darauf, ob ein Treatmenteffekt vorliegt und wie stark dieser ist, sondern auch darauf, ob es Gruppen gibt, für die ein besonders starker Treatmenteffekt vorliegt und andere, für die ein solcher nicht (oder gar unter umgekehrten Vorzeichen) zu erwarten ist (Bryan et al., 2021; VanderWeele et al., 2019). Gerade solche tiefgreifenderen Erkenntnisse liefern Aufschluss über die Wirkung kausaler Mechanismen und sind daher gerade für die Weiterentwicklung von Theorien integral. Als wie belastbar sind vor diesem Hintergrund die üblicherweise auf die Schätzung durchschnittlicher Effekte fokussierenden Befunde experimenteller Studien einzuschätzen?

1.2.3. Methodische Debatte um die „Künstlichkeit“ von Experimenten und daraus resultierende Probleme der Validität von Befunden

In der Methodenliteratur rund um sozialwissenschaftliche Experimente findet sich eine lange zurückreichende Debatte um die Frage nach der Validität von Befunden experimenteller Studien. Seit Campbell (1957) wird dabei häufig zwischen interner und externer Validität unterschieden (siehe auch Campbell und Stanley, 1966; Cook und Campbell, 1979). Interne Validität bezieht sich dabei auf den innerhalb des Experiments zu ziehenden Inferenzschluss und damit auf die „Kausalität“ eines beobachteten Zusammenhangs. Externe Validität zielt hingegen auf den Inferenzschluss außerhalb beziehungsweise über das Experiment hinaus und damit auf die „Generalisierbarkeit“ eines beobachteten Zusammenhangs. Trotz später noch feingliedriger vorgeschlagener Unterscheidungen (bspw. Cook und Campbell, 1979), wird weiterhin häufig auf die zwei grundlegenden Konzepte interner und externer Validität Bezug genommen (Degtiar und Rose, 2023; Findley et al., 2021).

Diese Konzeptualisierung von Validität und insbesondere das Verhältnis interner und externer Validität zueinander sind allerdings nicht unumstritten (Gërxhani und Miller, 2022). So wird gemeinhin angenommen, dass ein Zielkonflikt zwischen den beiden Konzepten bestehe und in der Forschungspraxis daher eine Abwägung erfolgen müsse, ob im Einzelfall der stärkeren Kontrolle der Erhebungssituation (und damit interner Validität) oder der Natürlichkeit der Erhebungssituation (und damit externer Validität) mehr Gewicht beigemessen werden sollte. Typischerweise geht damit die Vorstellung einher, dass sich Forschungsansätze auf einem Kontinuum zwischen solchen mit hoher interner (bei zugleich geringer externer) Validität (bspw. Laborexperimente) auf der einen Seite und solchen mit hoher externer (bei zugleich geringer interner) Validität (bspw. Beobachtungsdaten) auf der anderen Seite anordnen lassen. Zwischen diesen Polen ließen sich in Abstufungen weitere Ansätze wie beispielsweise Feld- oder Survey-Experimente verorten (Al-Ubaydli und List, 2015; Roe und Just, 2009).

Ziel eines theoriegeleiteten Vorgehens ist der empirische Test von Hypothesen, der Aufschluss über die Gültigkeit zugrundeliegender Mechanismen gibt und so zur Weiterentwicklung theoretisch fundierter Erkenntnisse beiträgt (bspw. Jackson und Cox, 2013). Dieses Leitmotiv zugrunde legend greift der weit verbreitete Einwand mangelnder Realitätsnähe, insbesondere von Laborexperimenten, zu kurz. In einer differenzierten Auseinandersetzung mit den typischerweise vorgebrachten Bedenken gegenüber Laborexperimenten argumentieren Falk und Heckman (2009) überzeugend, dass der Vorwurf zu hoher Künstlichkeit im Kern an der zu beantwortenden Frage vorbei gehe. Demzufolge ist einzig die kontrollierte Variation des Treatments ausschlaggebend für kausalanalytischen Erkenntnisgewinn und eben diese kontrollierte Variation ist gerade kennzeichnend für Laborexperimente. Dennoch stimmen in diesem Zusammenhang Befunde, die auf eine geringe externe Validität von Ergebnissen aus Laborexperimenten hindeuten, zurückhaltend (Bader et al., 2019). Andererseits scheinen selbst vergleichsweise abstrakte Stimuli (in Survey-Experimenten) zu replizierbaren Ergebnissen zu führen (Brutger et al., 2022). In jedem Fall unterstreichen diese Befunde die Bedeutung andauernder Replikation und Kreuzvalidierung von Ergebnissen − gerade auch mit komplementären Ansätzen und Samples. In diese Richtung weisen auch Forderungen nach einer breiteren Variation der untersuchten Kontexte und Populationen, um Befunde (labor-)experimenteller Forschung weiter abzusichern (Falk und Heckman, 2009).

Konzeptionell stellt sich in diesem Zusammenhang jedoch die Frage, inwiefern es

sich überhaupt um einen Zielkonflikt zwischen interner und externer Validität handelt und ob nicht vielmehr davon auszugehen ist, dass interne Validität eine grundlegende Voraussetzung für externe Validität darstellt (Jiménez-Buedo und Miller, 2010). In dieser Perspektive müsste also zunächst sichergestellt werden, dass ein Befund intern valide ist, bevor eine weitere Befassung mit externer Validität zielführend erscheint (Gërxhani und Miller, 2022). So verwundert auch nicht, dass ein großer Teil der Literatur zu kausaler Inferenz auf die interne Validität von Befunden fokussiert. Findley et al. (2021) stellen diese Fokussierung auf interne Validität hingegen in Frage. Sie sehen darin das Problem, dass mangelnde externe Validität zu ebenso gravierenden Fehlschlüssen führen könne wie mangelnde interne Validität. Letzten Endes tragen also nur diejenigen Befunde zu substanziellem Erkenntnisgewinn bei, die sowohl intern als auch extern valide sind. Nur dann liege eine unverzerrte Schätzung von Zusammenhängen in der Zielpopulation vor, wie etwa jüngst Degtiar und Rose (2023) betonen.

In der neueren Literatur wird zunehmend auch der Begriff der Generalisierbarkeit verwendet, um auf die externe Validität von Befunden abzuzielen (Lesko et al., 2017). Häufig wird dabei ferner auf die Transportabilität von Befunden rekurriert (Keiding und Louis, 2018; Pearl und Bareinboim, 2014), ohne das jedoch zu explizieren. Dabei würde eine konsequente Unterscheidung dieser beiden Konzepte zu einem hohen Maß analytischer Klarheit der vorgebrachten Argumente beitragen (Findley et al., 2021). Generalisierbarkeit meint sodann die Möglichkeit, von den Erkenntnissen aus einem Sample S_1 auf eine Population P_1 zu schließen, aus der dieses Sample gezogen wurde. Transportabilität hingegen bezeichnet den Inferenzschluss von besagtem Sample S_1 auf eine andere Population P_2, aus der S_1 nicht stammt. Um die Belastbarkeit von Befunden valide einschätzen zu können, scheint es daher geboten, den Unterschied zwischen der Generalisierbarkeit der beobachteten Befunde (für die Population, aus der sie stammen) und der Transportabilität der beobachteten Befunde (auf andere Zielpopulation(en), aus der sie nicht stammen) zu beachten (Findley et al., 2021).

Eine vielversprechende Option zum Umgang mit dem skizzierten Spannungsfeld zwischen interner und externer Validität besteht etwa darin, durch die gezielte Kombination unterschiedlicher Ansätze sowohl die interne als auch die externe Validität zu erhöhen (Gërxhani und Miller, 2022). Zu denken ist hier beispielsweise an Feldexperimente, faktorielle Survey-Experimente und dergleichen mehr. Diese liefern in zumindest zweierlei Hinsicht wichtige Beiträge. Erstens erlauben sie die empirische Überprüfung an heterogenen Samples aus der zugrunde liegenden Inferenzpopulation

und damit auch einen Test des Treatmenteffekts unter einer umfassenderen Variation des äußeren Kontexts (Falk und Heckman, 2009). Zweitens bieten sie vielfältige Möglichkeiten für Replikationen, die (sofern sie zu substanziell übereinstimmenden Ergebnissen führen) das Vertrauen in die ursprünglich zu testenden theoretischen Überlegungen stärken (Jackson und Cox, 2013). Wie ist jedoch mit nicht oder nur zum Teil übereinstimmenden Ergebnissen von Replikationsbestrebungen umzugehen? Diese könnten Anlass dazu geben, die ursprünglich berichteten Befunde in Zweifel zu ziehen oder aber Anstoß dazu liefern, genauer zu hinterfragen, ob unter diesen Umständen tatsächlich mit dem erwarteten Effekt zu rechnen gewesen ist, was wiederum auf eine oftmals erforderliche systematischere Auseinandersetzung mit Heterogenität verweist.

1.2.4. Effektheterogenität als zentrale Herausforderung

Unbenommen erweisen sich die verschiedenen Varianten sozialwissenschaftlicher Experimente als besonders geeignet für den Test kausaler Hypothesen. Wie in den vorigen Abschnitten dargelegt, erlauben sie eine unverzerrte Schätzung des zu untersuchenden Treatmenteffekts. Qua randomisierter Zuweisung auf Treatment- und Kontrollgruppe wird sichergestellt, dass keine anderen Einflüsse als die des Treatments selbst einen Unterschied zwischen Treatment- und Kontrollgruppe erklären können (Jackson und Cox, 2013).

Unter der Annahme homogener Treatmenteffekte ist dann zu erwarten, dass es keine Rolle spielt, nach welchen Kriterien Proband:innen für die Analyse rekrutiert werden. Folglich wird davon ausgegangen, dass sich Treatmenteffekte in unterschiedlichen Kontexten und mit unterschiedlichen Samples replizieren lassen sollten. Liegen allerdings tatsächlich heterogene Treatmenteffekte vor, wie es wohl für einen Großteil sozialwissenschaftlicher Fragestellungen zu erwarten ist, gestaltet sich die korrekte Identifikation schwieriger (Kohler et al., 2019).

Eine zu große Gewissheit ausgerechnet um die methodischen Stärken experimenteller Designs bei der Schätzung kausaler Zusammenhänge könnte an der Stelle insofern ein Hindernis darstellen, als dass sie dazu verleiten könnte, den Blick nicht auf mögliche Verzerrungen durch Heterogenität zu richten. So könnte ein beispielhaft genannter Treatmenteffekt trotz experimentellen Designs etwa nur unter bestimmten Bedingungen oder für eine spezifische Population zu erwarten sein. Wird das in der Konzeption einer empirischen Studie fälschlicherweise übersehen, kann kein

belastbarer Schluss auf eine breiter definierte Inferenzpopulation erfolgen. Vielmehr ist damit zu rechnen, dass Versuche, einen solchen Effekt zu replizieren, scheitern oder zumindest zu einem uneinheitlichen Forschungsstand beitragen.

Empirisch lässt sich das etwa an der bereits angerissenen Debatte um die Wirksamkeit von (grünen) Nudges zeigen. Bezug nehmend auf die initial vielversprechenden Vorschläge von Thaler und Sunstein (2021) befasste sich eine Vielzahl empirischer Untersuchungen mit der Wirksamkeit unterschiedlicher Nudging-Interventionen. Die teils vielbeachteten Beispiele (zu denken ist etwa an die Opower energy conservation programs) deuteten auf zunächst sehr vielversprechende Befunde hin, die im weiteren Verlauf jedoch nicht repliziert werden konnten beziehungsweise deutlich geringere Effektstärken aufwiesen (Allcott, 2011, 2015). Dieser Eindruck einer unklaren Befundlage wird jenseits der Betrachtung einzelner Anwendungsbeispiele etwa durch die kürzlich erschienene Metastudie von Mertens et al. (2022a,b) verstärkt. Sie finden basierend auf mehr als 200 Studien zu unterschiedlichen Nudging-Interventionen nur kleine bis moderate Effekte, aber auch Anhaltspunkte für einen Publikationsbias hin zu mehr positiven Effekten in der Literatur. Die in Reaktion darauf angeregten Beiträge zeigen, dass die berichteten Treatmenteffekte noch überschätzt sind und unter Verwendung alternativer Modellierungsverfahren zur Korrektur für den Publikationsbias vollständig verschwinden (Bakdash und Marusich, 2022; Maier et al., 2022). Szaszi et al. (2022) betonen darüber hinaus, dass ein klarerer Fokus auf die Heterogenität von Effekten und dafür verantwortliche Ursachen gelegt werden sollte, anstatt auf durchschnittliche Effekte abzustellen (Mertens et al., 2022c; Szaszi et al., 2022). Ein solcher Fokus erlaubt es schließlich, mögliche Erklärungen für Effektheterogenität zu finden, worauf beispielsweise die Befunde von Berger et al. (2022) zu zwar großen, mit steigenden Kosten allerdings abnehmenden Effektstärken für grüne Default-Optionen beim Flugticket-Kauf hindeuten.

Das knüpft an die Argumentation von Bryan et al. (2021) an, die eine systematische Auseinandersetzung mit Heterogenität fordern. Vor dem Hintergrund einer sich über Teile verschiedener (auch sozialwissenschaftlicher) Disziplinen erstreckenden Replikationskrise (Auspurg und Brüderl, 2021, 2022) sprechen sie sich für einen Paradigmenwechsel aus. So könne nur durch eine differenzierte Befassung mit der Frage, für wen unter welchen Bedingungen und warum ein bestimmter Effekt überhaupt zu erwarten ist, ein tiefergreifendes Verständnis der ursächlichen Mechanismen vorangetrieben werden. Um dabei auf theoretisch potenziell relevante Moderatoren testen zu können, müssen allerdings auch entsprechende Datenpunkte vorhanden

sein, wozu diese bereits in der Konzeptionsphase von Studien berücksichtigt werden müssen. Der in der vorliegenden Arbeit vorgeschlagene, gezielt theoriegeleitete Einbezug von Kontextinformationen stellt darauf ab, an dieser Stelle einen Beitrag zu leisten.

1.3. Der Beitrag einer (stärkeren) Kombination mit Kontextinformationen

Anliegen der vorliegenden Dissertation ist es, Möglichkeiten aufzuzeigen, wie der Einbezug von Kontextinformationen dazu genutzt werden kann, Effektheterogenität im sozialwissenschaftlichen Experiment zu berücksichtigen. Bevor im Folgenden näher auf die einzelnen Beiträge hinsichtlich ihrer verbindenden methodischen Fragestellung eingegangen wird, soll zunächst das methodische Vorgehen der vorliegenden Arbeit erläutert werden. Die Ausführungen sind dabei an das fundierte Plädoyer von Bryan et al. (2021) für eine stärkere Fokussierung auf Heterogenität angelehnt.

Wesentlicher Kern des Ansatzes ist die Einsicht, dass für ein tiefgreifendes Verständnis zugrundeliegender Mechanismen eine theoretisch fundierte empirische Auseinandersetzung mit möglichen Moderatoren erfolgen muss. So ist jenseits der Frage, ob ein bestimmtes Treatment einen kausalen Effekt entfaltet, zu hinterfragen, für welche Proband:innen ein solcher Effekt unter welchen Bedingungen und in welcher Stärke tatsächlich zu erwarten ist. Diese Überlegungen verweisen letztlich darauf, dass es weit differenzierter zu klären gilt, warum und wie genau ein Treatment wirkt. Eine weitgehende Fokussierung auf durchschnittliche Effekte hingegen läuft Gefahr, derlei wesentliche Aspekte zugrunde liegender Mechanismen und deren Bedeutung für die Weiterentwicklung theoretischer Überlegungen zu stark auszublenden.

Welchen Mehrwert die explizite Modellierung theoretisch zu erwartender Heterogenität über den Wertebereich eines Moderators hinweg bietet, lässt sich anhand der graphischen Visualisierung eines stilisierten Beispiels erläutern. So zeigt Abbildung 1.1 die Verteilung theoretischer Treatmenteffekte individueller Proband:innen einer hypothetischen Inferenzpopulation (vgl. Bryan et al., 2021, Abb. 1). In Panel (a) sind die theoretischen Treatmenteffekte in einem Streudiagramm über den Wertebereich eines Moderators hinweg dargestellt. Panel (b) hingegen zeigt einen Stripplot derselben theoretischen Treatmenteffekte ohne Berücksichtigung des Moderators. Wie die stilisierte Darstellung zeigt, würde man ohne Berücksichtigung

des Moderators bestenfalls eine durchschnittliche Schätzung des Treatmenteffekts in der gesamten Population erzielen können, die jedoch wesentliche Heterogenität ausblendet. Ein genauerer Blick auf Panel (a) offenbart darüber hinaus, dass auch die Schätzung eines durchschnittlichen Treatmenteffekts in hohem Maße davon abhängt, wie ein hypothetisches Sample schlussendlich zusammengesetzt ist. Während beispielsweise eine hypothetische Studie A die gesamte Bandbreite des Moderators abzudecken vermag, enthält eine hypothetische Studie B qua Design lediglich Proband:innen mit bestimmten Ausprägungen des Moderators, wohingegen eine weitere hypothetische Studie C wiederum auf ein anderes Segment des Moderators beschränkt bleibt. Es ist offenkundig, dass die drei hypothetischen Studien in diesem Beispiel zu substanziell unterschiedlichen Schätzungen des durchschnittlichen Treatmenteffekts kommen. So ginge man basierend auf den Ergebnissen der hypothetischen Studie A von einer (fälschlicherweise für alle Proband:innen gleichermaßen zu erwartenden) moderaten Effektstärke aus, wohingegen man anhand von Studie B einen stark ausgeprägt positiven Effekt vermuten würde, den man basierend auf Studie C überhaupt nicht nachweisen könnte.

Um also jenseits des stilisierten Beispiels tatsächlich in die Lage versetzt zu sein, Heterogenität über den Werteberich eines potenziellen Moderators hinweg prüfen oder gar explizit modellieren zu können, müssen in Frage kommende Moderatoren zuvor (theoretisch) identifiziert und (empirisch) gemessen werden. Bereits in der Konzeptionsphase eines Experiments muss also bedacht werden, welche Moderationen für die jeweilige Fragestellung relevant erscheinen könnten, um diese einer späteren Analyse zugänglich machen zu können. Zu betonen ist in diesem Zusammenhang, dass auch nicht experimentell variierte beziehungsweise nicht experimentell variierbare Variablen moderierend wirken könnten (bspw. Einstellungen der Proband:innen gegenüber latenten Konstrukten, wie etwa Umweltbewusstsein). Dasselbe trifft auf (räumliche und zeitliche) Bedingungen zu, unter denen ein Experiment durchgeführt wird (bspw. lokale Marktsituationen, in denen ein mögliches Treatment auf unterschiedlich ausgeprägte Reaktionsbereitschaft im Sample trifft).

Zweifellos lässt sich das beschriebene Vorgehen auch anwenden, wenn angenommen wird, dass die moderierend wirkende Variable nicht kontinuierlich, sondern diskret ist. Entsprechend zeigt Abbildung 1.2 wiederum die Verteilung theoretischer Treatmenteffekte individueller Proband:innen einer hypothetischen Inferenzpopulation analog zum gerade beschriebenen Vorgehen. In Panel (a) sind Stripplots der theoretischen Treatmenteffekte verschiedener Subgruppen dargestellt. Panel (b)

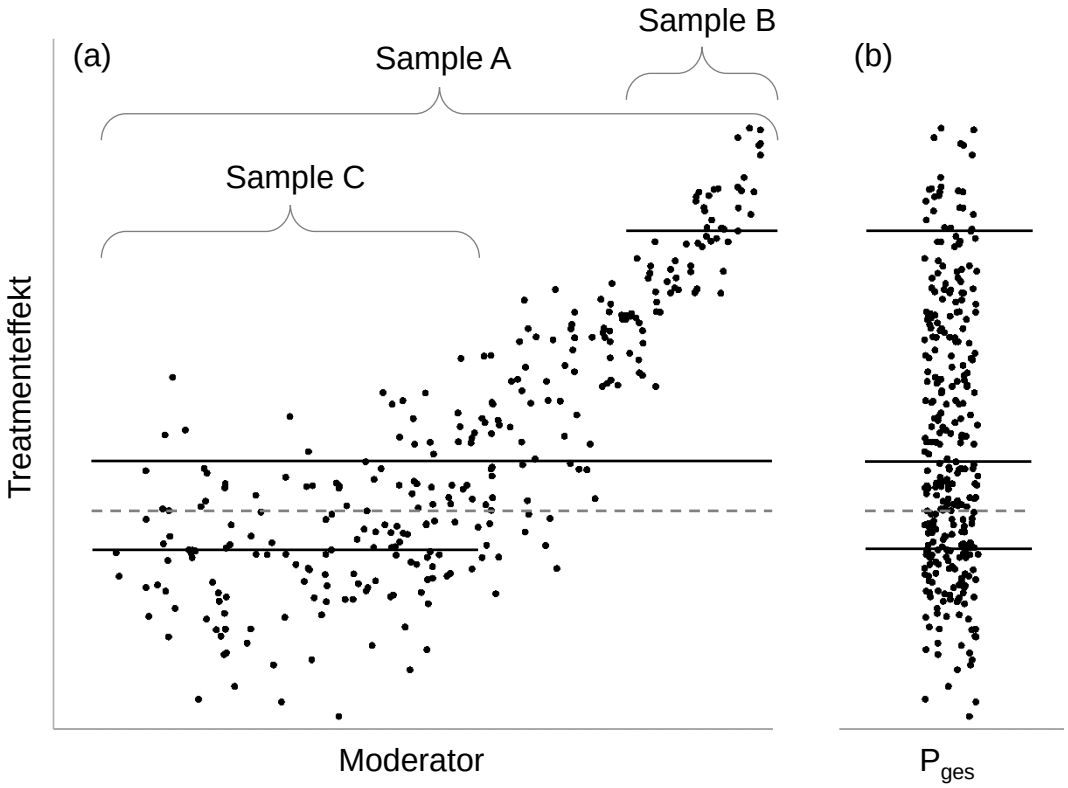

Abbildung 1.1.: Schematische Darstellung der theoretischen Treatmenteffekte individueller Proband:innen einer hypothetischen Population mit und ohne Berücksichtigung eines Moderators, angelehnt an Bryan et al. (2021, Abb. 1)

Anmerkungen: Das Streudiagramm in Panel (a) zeigt die Verteilung der Treatmenteffekte über den Wertebereich eines Moderators hinweg; in Panel (b) sind dieselben Treatmenteffekte ohne Berücksichtigung des Moderators dargestellt. Punkte stehen dabei für den theoretischen Treatmenteffekt individueller Proband:innen. Die geschweiften Klammern deuten auf verschiedene hypothetische Samples A-C, die aus der Population gezogen werden könnten. Die zu den verschiedenen hypothetischen Samples jeweils korrespondierenden durchschnittlichen Treatmenteffekte sind als durchgezogene Linien dargestellt. Die gestrichelte Referenzlinie deutet auf einen Nulleffekt.

zeigt einen zusammengefassten Stripplot derselben theoretischen Treatmenteffekte ohne Berücksichtigung der Zugehörigkeit zu einer der hypothetischen Subgruppen. Wie die stilisierte Darstellung zeigt, würde man je nachdem, welche der Subgruppen in einer Studie untersucht wird, zu grundverschiedenen Einsichten gelangen. So würde in diesem Beispiel eine Schätzung des durchschnittlichen Treatmenteffekts für Gruppe A negativ ausfallen, für die Gruppen B und C wäre im Schnitt von ei-

nem Nulleffekt auszugehen, wohingegen in Gruppe D ein im Schnitt positiver Effekt vorläge. Bei ausschließlicher Betrachtung des durchschnittlichen Effekts insgesamt würden aufschlussreiche Unterschiede zwischen den Subgruppen übersehen. Aber auch umgekehrt, wenn eine angestrebte Analyse im Kern auf die Interpretation nur einzelner Subgruppen ausgerichtet ist, blieben die teils entgegengesetzten Befunde anderer Subgruppen verdeckt und die Schätzung eines insgesamt zu erwartenden durchschnittlichen Effekts im gesamten Sample wäre verzerrt.

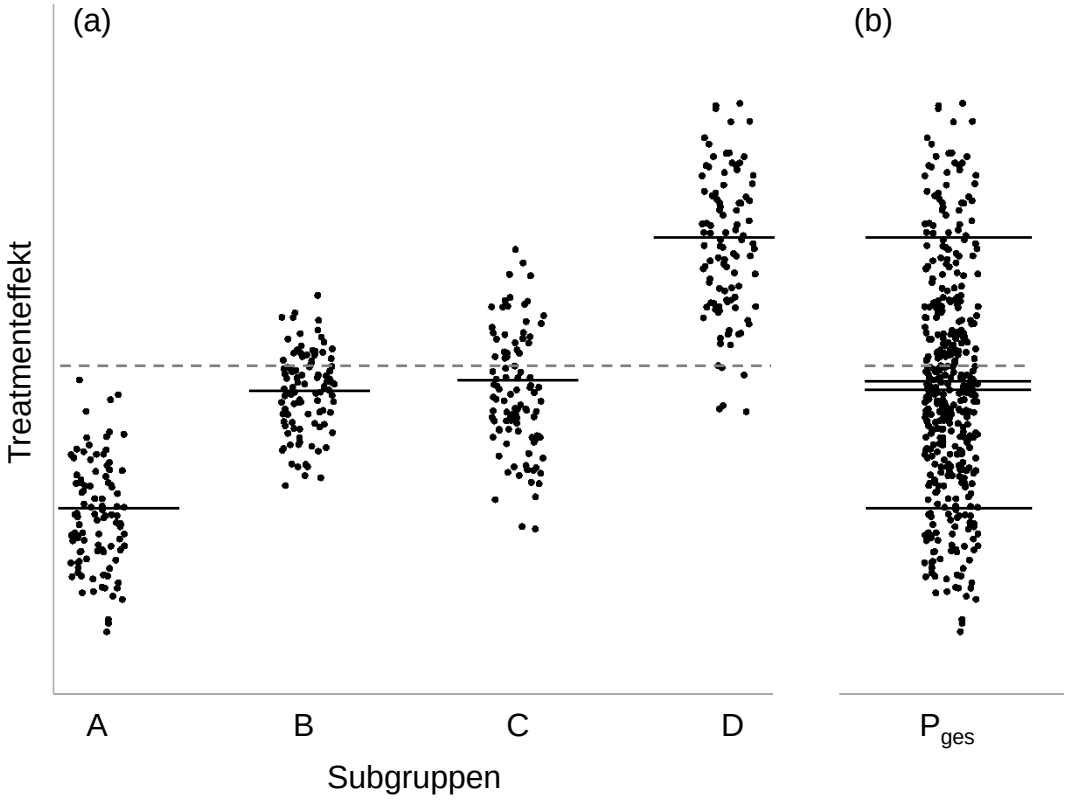

Abbildung 1.2.: Schematische Darstellung der theoretischen Treatmenteffekte individueller Proband:innen einer hypothetischen Population mit und ohne Berücksichtigung verschiedener Subgruppen

Anmerkungen: Der Stripplot in Panel (a) zeigt die Verteilung der Treatmenteffekte für hypothetische Subgruppen A-D; in Panel (b) sind dieselben Treatmenteffekte ohne Unterscheidung nach Subgruppe dargestellt. Punkte stehen dabei für den theoretischen Treatmenteffekt individueller Proband:innen. Die zu den verschiedenen Subgruppen beziehungsweise der Population insgesamt jeweils korrespondierenden durchschnittlichen Treatmenteffekte sind als durchgezogene Linien dargestellt.Die gestrichelte Referenzlinie deutet auf einen Nulleffekt.

Um nun diese unterschiedlichen Subgruppen identifizieren und für die empirische Messung berücksichtigen zu können, scheint es abermals sinnvoll, weitere Informationen mit den Beobachtungen im Experiment zu verknüpfen. Dabei kommen je nach Fragestellung unterschiedliche Variablen zur Gruppierung in sich homogener Subgruppen infrage (bspw. ließe der räumliche Kontext, in dem ein Treatment erfolgt Rückschlüsse darauf zu, inwiefern Proband:innen selbst stärker oder schwächer von einem Treatment betroffen sind und mithin unterschiedliche Reaktionen zu erwarten sein könnten).

Den verschiedenen Varianten von Effektheterogenität könnte eine ausschlaggebende Rolle zukommen, wie die nachfolgend zu diskutierenden empirischen Anwendungsbeispiele veranschaulichen sollen. Sie sind bewusst so gewählt, dass sich an einer Bandbreite verschiedener Forschungsfragen aufzeigen lässt, wie der Einbezug von Kontextinformationen im sozialwissenschaftlichen Experiment produktiv erfolgen kann. Nach einer kurzen Verortung der einzelnen Beiträge befasst sich das erste Anwendungsbeispiel mit der Modellierung von theoretisch erwarteter Heterogenität über den Wertebereich eines Moderators hinweg. Fokussiert wird dabei auf die funktionale Form eines nicht-linearen Zusammenhangs. Im zweiten Anwendungsbeispiel werden Kontextfaktoren genutzt, um relevante Subgruppen zu identifizieren und so Heterogenität durch die getrennte Schätzung von Treatmenteffekten zu verringern. Im dritten Anwendungsbeispiel wird der Einbezug von Kontextinformationen dazu genutzt, Heterogenität in der Zusammensetzung des Samples zu identifizieren, wobei das Potential der vorgeschlagenen Kombination in der gezielten Kontrolle der Feldbedingungen liegt.

1.3.1. Zusammenfassung und Einordnung der eigenen Beiträge

Die vorliegende Dissertation setzt sich im Kern aus drei bereits publizierten Beiträgen zusammen. Dabei handelt es sich um zwei in Fachzeitschriften erschienene Artikel und ein Kapitel in einem Forschungshandbuch (siehe Tabelle 1.1). Die Beiträge sind sowohl in der deutschsprachigen Soziologie (Thiel, 2020, ; siehe Kapitel 2) als auch in internationalen Publikationsorganen (Thiel, 2021, ; siehe Kapitel 3), zum Teil zusammen mit Koautor:innen (Auspurg et al., 2020, ; siehe Kapitel 4), erschienen.

Die den Kapiteln 2 und 3 zugrundeliegenden Daten wurden im Rahmen eines Lehrforschungsprojekts an der LMU München erhoben. Die Konzeption der Erhe-

Tabelle 1.1.: Übersicht der eigenen Beiträge

Kapitel	Publiziert in	Impact Factor[d]	Koautor:innen	Fremdanteil[e]	Eigenanteil	Gewichtung[f]	Score[g]
2	KZfSS[a]	1,526 (2020)			100%	1,5	1,5
3	Handbook Env. Soc.[b]				100%	1,25	1,25
4	RSSM[c]	1,604 (2020)	Katrin Auspurg, Andreas Schneck	40%, 20%	40%	1,5	0,6
Σ							3,35

Anmerkungen: [a]Kölner Zeitschrift für Soziologie und Sozialpsychologie; [b]Research Handbook on Environmental Sociology; [c]Research in Social Stratification and Mobility; [d]Scopus 3-Jahres Impact Faktoren im jeweiligen Erscheinungsjahr (in Klammern); [e]In der Reihung der Koautor:innen; [f]Gewichtung nach Betreuungsvereinbarung; [g]Ergibt sich aus Eigenanteil und Gewichtung der Beiträge, in Summe muss ein Score > 3 erreicht werden.

bung und des darin enthaltenen faktoriellen Survey-Experiments erfolgte in enger Zusammenarbeit mit Sabine Düval und Katrin Auspurg. Alle darüber hinaus gehenden Schritte der Datenaufbereitung und -auswertung für die in den Kapiteln 2 und 3 berichteten Studien sind allein dem Autor zuzurechnen.

Die Datenbasis für Kapitel 4 entstammt einem im Rahmen des Projekts „Ethnische Benachteiligung und Segregation in Wohnungsmärkten“ durchgeführten Feldexperiment. Die Konzeption und Durchführung des Experiments sowie die Erfassung der Daten zum Wohnungsmarkt ist dabei insbesondere Katrin Auspurg und Andreas Schneck zu verdanken. Die umfassende Aufbereitung der Marktdaten für die hier untersuchte Fragestellung erfolgte maßgeblich durch den Autor selbst. Ein erster Textentwurf insbesondere der theoretischen Rahmung und breiteren Einbettung des Beitrags erfolgte durch Katrin Auspurg. Die übersichtliche Aufbereitung des Forschungsstandes ist Andreas Schneck zu verdanken. Die Analyse der Daten und die Aufbereitung der Ergebnisse sowie die im Methodenanhang berichteten Simulationen unterschiedlicher Stichprobenverfahren gehen auf den Autor selbst zurück. Die Diskussion der Befunde, die finale Ausarbeitung des Manuskripts sowie sämtliche Überarbeitungen erfolgten in enger Absprache der Autor:innen und lassen sich kaum einzeln zurechnen. Eine Übersicht der insgesamt geleisteten Arbeitsanteile ist Tabelle 1.1 zu entnehmen.

1.3.2. Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut

Im ersten Beitrag (Kapitel 2) wird die in der Einstellungsforschung lange zurückreichende Frage aufgegriffen, warum sich Einstellungen oftmals nicht in entsprechendem Verhalten niederschlagen. So ist in verschiedenen Bereichen menschlichen Handelns eine vielfach nur mäßige Korrelation zwischen Einstellung und Verhalten zu beobachten. Aus umweltsoziologischer Perspektive stellt sich mithin die Frage, wie es zu erklären ist, dass beispielsweise ein hohes Umweltbewusstsein oftmals nicht zu umweltverträglicheren Verhaltensentscheidungen führt. Ein in diesem Zusammenhang viel diskutierter Ansatz ist die sogenannte Low-Cost-Hypothese (LCH). Sie besagt im Kern, dass der Einfluss einer Einstellung neben der Stärke der Einstellung selbst auch davon abhänge, wie kostspielig einstellungskonformes Verhalten ausfällt. Die Idee ist also, dass mit steigender Kostspieligkeit einstellungskonformen Verhaltens der Einfluss einer Einstellung auf die Verhaltenswahl sinkt. Im Anschluss an grundlegende Arbeiten von Diekmann und Preisendörfer (1998, 2003) wurde die LCH vielfach aufgegriffen. Sie wurde an einer Reihe unterschiedlicher Anwendungsbeispiele empirisch getestet und ist darüber hinaus auch immer wieder Gegenstand intensiver theoretischer Diskussion (bspw. Best und Kroneberg, 2012; Keuschnigg und Kratz, 2018; Tutić et al., 2017).

Trotz der umfangreichen Rezeption hat die Literatur um die LCH bislang eine unklare Befundlage mit teils widersprüchlichen Ergebnissen hervorgebracht. Eine mögliche Ursache für diesen Umstand könnte in der konkreten Modellierung der durch die LCH postulierten Zusammenhänge liegen. So sieht die gängige Teststrategie vor, neben den Haupteffekten der Einstellung und der Kosten einstellungskonformen Verhaltens auch einen Interaktionsterm aus den beiden aufzunehmen. Unter Rückgriff auf ein mikroökonomisches Nachfragemodell schlagen Tutić et al. (2017) hingegen vor, auf die explizite Aufnahme eines solchen Interaktionsterms zu verzichten. Stattdessen schlagen sie eine Teststrategie vor, die das Zusammenspiel aus Einstellung und Kosten implizit abbildet. Der Test der LCH im Rahmen einer linearen Regressionsanalyse könne mithin anhand der logarithmierten Nachfragefunktion nach einstellungskonformem Verhalten erfolgen. Diese setzt sich additiv aus der Stärke der Einstellung, dem verfügbaren Einkommen sowie dem Preis einstellungskonformen Verhaltens (in jeweils logarithmierter Form) zusammen.[9] Tutić et al. (2017)

9 Wobei zu berücksichtigen ist, dass „Einkommen“ und „Preis“ hier in einem weiten Sinn zu

veranschaulichen diese neue Teststrategie anhand zweier empirischer Beispiele, operationalisieren die Kostspieligkeit einstellungskonformen Verhaltens dabei jedoch lediglich über die Begrenztheit des verfügbaren Einkommens. So testen sie eine weite Interpretation der LCH im Sinne eines Einkommenseffekts (und sprechen in diesem Zusammenhang auch von einer Low-Income-Hypothese). Ein Test im engeren Sinn eines Preiseffekts einstellungskonformen Handelns, wie es die von ihnen vorgeschlagene Modellierung vorsähe, lässt diese Operationalisierung hingegen nicht zu.

Am Beispiel der Befürwortung einer City-Maut zur Verbesserung der Luftqualität in größeren Städten liefert der vorliegende Beitrag eine erste empirische Überprüfung der LCH im engeren Sinn eines Preiseffekts anhand der von Tutić et al. (2017) vorgeschlagenen Modellierung. Datengrundlage bildet eine im Frühsommer 2018 in München und einigen Umlandgemeinden mit hoher Pendelverflechtung zu München durchgeführte postalische Bevölkerungsbefragung. Kern der Erhebung war ein in den Rahmenfragebogen integriertes faktorielles Survey-Experiment zur Befürwortung einer möglichen City-Maut in München. Den Befragten wurden dabei jeweils 4 Beschreibungen (Vignetten) fiktiver Mautmodelle zur Bewertung vorgelegt. Die konkrete Ausgestaltung der Merkmale (Dimensionen) der verschiedenen Mautmodelle, wie beispielsweise die Höhe der vorgesehenen Mautgebühren aber auch der Grad der voraussichtlich zu erwartenden Luftverbesserungen im Stadtgebiet, wurde experimentell variiert. Insgesamt wurden über 1300 Befragten mehr als 5300 Vignetten zur Bewertung vorgelegt.

Die durch die LCH postulierten Zusammenhänge lassen sich durch eine Kombination des faktoriellen Survey-Experiments mit weiteren Angaben der Befragten empirisch überprüfen. Die Befürwortung einer City-Maut (beziehungsweise genau genommen die Absicht, eine solche zu befürworten), dient dabei als Maß umweltgerechten Verhaltens. Bei gegebenem Einkommen (gemessen durch das Haushaltsnettoäquivalenzeinkommen) ergibt sich durch die experimentelle Variation der Höhe der vorgesehenen Mautgebühren und damit des Preises umweltgerechten Verhaltens eine exogene Variation der Kostspieligkeit. Dieses Vorgehen stellt einen wesentlichen Vorteil gegenüber der rein endogen erfolgenden Operationalisierung der Kostspieligkeit anhand des Einkommens dar und erlaubt zudem eine differenziertere Analyse der postulierten Zusammenhänge. Als Maß der Stärke der Einstellung dient das mithilfe einer Item-Batterie erfasste Umweltbewusstsein der Befragten. Die zu

verstehen sind und jenseits der monetären Dimension auch beispielsweise auf verfügbare Zeit-Budgets und Opportunitätskosten abstellen können (Tutić et al., 2017).

testende Vermutung ist, dass der Einfluss der Kostspieligkeit der Befürwortung einer City-Maut über den Wertebereich des Umweltbewusstseins hinweg variiert, der Einfluss der Einstellung (des Umweltbewusstseins) also mit steigender Kostspieligkeit einstellungskonformen (umweltgerechten) Verhaltens abnimmt. Die verbesserte Teststrategie erlaubt es, diese theoretisch zu erwartende Heterogenität empirisch zu überprüfen. Es zeigt sich, dass die Zusammenhänge tatsächlich weitgehend der theoretisch erwarteten funktionalen Form folgen. Die Ergebnisse erweisen sich darüber hinaus als sehr robust.[10] Der Beitrag knüpft damit in mehrerlei Hinsicht an die bestehende Forschung an. So liefert er einen ersten Test der LCH im engeren Sinn eines Preiseffekts einstellungskonformen Handelns und trägt so zur umfangreichen methodischen Diskussion um die LCH bei. Zudem abstrahiert der Beitrag von einer bloßen Replikation der vorgeschlagenen Modellierung, indem die Überlegungen an einem anderen, inhaltlich ebenso aufschlussreichen Anwendungsfall getestet werden. Für die Entwicklung umwelt- und klimaschonenderer Mobilitätskonzepte ist etwa im Hinblick auf Ballungsräume neben Fragen der individuellen Verkehrsmittelwahl interessant, inwiefern Bereitschaft zur Gestaltung der vorgelagerten Rahmenbedingungen besteht, innerhalb derer individuelle Mobilitätsentscheidungen getroffen werden und welche Rolle Umweltbewusstsein in diesem Zusammenhang spielt.

1.3.3. Support for city road tolls: a question of self-interest?

Im zweiten Beitrag der Dissertation (Kapitel 3) wird die Frage der Befürwortung einer City-Maut nochmals allgemeiner, jenseits der spezifischen Überlegungen rund um die Low-Cost-Hypothese, aufgegriffen. Der Beitrag zielt darauf ab, zu einem genaueren Verständnis der Faktoren beizutragen, die für die Befürwortung einer City-Maut maßgeblich sind. Im Fokus steht dabei die Frage, ob die insgesamt zu beobachtenden geringen Zustimmungswerte tatsächlich für die gesamte Bevölkerung gleichermaßen zu erwarten sind oder ob hier nicht vielmehr mit substanzieller Heterogenität zu rechnen ist.

Während sich eine Vielzahl von Studien der Frage verschrieben hat, den Einfluss unterschiedlicher Aspekte auf die Befürwortung einer City-Maut zu untersuchen,

10 So wurden etwa im Rahmen weiterer Robustheitsanalysen Angaben zur bisherigen Autonutzung der Befragten im Münchner Stadtgebiet dazu genutzt, die Höhe der individuell tatsächlich zu erwartenden Mautgebühren zu berechnen. Auch für die Schätzung unter Verwendung von diesem Maß, das zu erwartende Kostenunterschiede zwischen häufiger und seltener Autonutzung abbildet, bleiben die Ergebnisse substanziell robust.

blenden diese Studien, von einzelnen Ausnahmen abgesehen, typischerweise mögliche Unterschiede zwischen verschiedenen Subgruppen aus. Zu nennen ist in diesem Zusammenhang etwa eine Untersuchung zur Befürwortung einer City-Maut in London und Leeds, bei der Jaensirisak et al. (2005) nach der Autonutzung der Befragten differenzieren und erwartungsgemäß neben höherer Zustimmung für nur geringe Mautgebühren auch höhere Zustimmungswerte unter Nicht-Autofahrer:innen finden. Hinsichtlich einer Differenzierung nach Anwohnerschaft ist die Befundlage weit unklarer. So finden etwa Rentziou et al. (2011) unter Anwohner:innen in Athen eine geringere Unterstützung für eine mögliche City-Maut, wohingegen Milenković et al. (2019) unter Anwohner:innen in Belgrad höhere Zustimmungswerte zu einer möglichen City-Maut finden als unter Personen, die außerhalb des betroffenen innerstädtischen Gebiets wohnen. Gegenüber der bereits bestehenden City-Maut in Stockholm finden Eliasson und Jonsson (2011) schließlich negativere Bewertungen unter Anwohner:innen, allerdings nur unter Verwendung bestimmter Modellspezifikationen. Ein systematischer Fokus auf diese möglichen Ursachen substanziell unterschiedlicher Haltungen gegenüber einer City-Maut könnte zu einem tiefgreifenderen Verständnis beitragen.

Die Idee des vorliegenden Beitrags ist es, sich die Kombination eines faktoriellen Survey-Experiments mit zusätzlichen Angaben zum Mobilitätsverhalten und Informationen zur Wohnumgebung der Befragten zunutze zu machen, um die der Befürwortung einer City-Maut zugrundeliegenden Mechanismen zu entflechten. Leitend ist dabei die Frage, inwiefern mit substanziellen Unterschieden zwischen Subgruppen zu rechnen ist, die in unterschiedlichem Maße von den bisherigen Problemen innerstädtischer Verkehrsüberlastung aber auch von konkreten Gegenmaßnahmen wie beispielsweise einer möglichen City-Maut betroffen wären. Dazu wird erneut auf die Daten der im Frühsommer 2018 in München und einigen Umlandgemeinden durchgeführten Bevölkerungsbefragung zurückgegriffen, die bereits in Kapitel 2 genutzt wurde. Dabei erfolgen zunächst Schätzungen der Befürwortung fiktiver Ausgestaltungen einer möglichen City-Maut für das gesamte Sample. In einem weiteren Schritt werden Kontextinformationen gezielt dazu genutzt, verschiedene Subgruppen zu identifizieren, die von unterschiedlichen Motivlagen gegenüber einer möglichen City-Maut geprägt sein dürften. Dabei wird in Erweiterung bestehender Ansätze die Unterscheidung nach Anwohnerschaft (ja/nein) und Autonutzung (ja/nein) gekreuzt betrachtet und Analysen entsprechend auch für die so gebildeten Subgruppen getrennt durchgeführt.

Empirisch zeigt sich anhand des gesamten Samples in Übereinstimmung mit der Literatur wenig Befürwortung einer möglichen City-Maut. Neben dieser insgesamt kritischen Haltung sind allerdings substanzielle Unterschiede zwischen Subgruppen zu beobachten. So zeigt sich, dass insbesondere Autofahrer:innen von Außerhalb starke Ablehnung einer möglichen City-Maut äußern. Befragte ohne Auto, die innerhalb des Geltungsbereichs einer fiktiven City-Maut wohnen, stehen einer solchen hingegen tendenziell aufgeschlossen gegenüber. Sowohl Autofahrer:innen, die innerhalb des Geltungsbereichs einer fiktiven City-Maut wohnen als auch Befragte ohne Auto, die außerhalb wohnen, liegen mit ihrer Haltung gegenüber einer möglichen City-Maut erwartungsgemäß zwischen diesen beiden Polen. Dieses Muster deutet auf substanzielle Heterogenität, die ohne den Einbezug von Kontextinformationen typischerweise übersehen wird.

Ein genauerer Blick auf die unterschiedliche Bedeutung einzelner Dimensionen der fiktiven zur Bewertung vorgelegten Ausgestaltungen einer City-Maut deutet zudem auf unterschiedliche Erklärungen für diesen Befund hin. So könnte etwa eine Ausnahmeregelung für Anwohner:innen einerseits aus allgemeinen „Fairness“-Überlegungen heraus befürwortet werden. Immerhin könnte diese eine Kompensation für die ohnehin hohen Belastungen in Folge des hohen innerstädtischen Verkehrsaufkommens darstellen (bspw. Huber et al., 2020). Andererseits könnte die Befürwortung einer solchen Befreiung von vorgesehenen Mautgebühren gerade im genuinen Eigeninteresse von Anwohner:innen liegen, die selbst über ein Auto verfügen (bspw. Rohrschneider, 1988). Empirisch zeigt sich, dass vor allem Autofahrer:innen, die innerhalb des Geltungsbereichs einer möglichen City-Maut wohnen, solche Ausgestaltungen deutlich positiver bewerten, die eine Befreiung für Anwohner:innen vorsehen als solche Vorschläge, die keine Befreiung vorsehen. Aber auch Autofahrer:innen von Außerhalb bewerten diese (etwas) positiver, wohingegen diese Dimension für Nicht-Autofahrer:innen, ganz gleich ob innerhalb oder außerhalb des Geltungsbereichs einer möglichen City-Maut wohnend, keine Bedeutung hat. Neben allgemeinen Überzeugungen hinsichtlich der Implikationen solcher Maßnahmen erweist sich insbesondere auch das individuelle Eigeninteresse als relevant. Insofern leistet die differenzierte Analyse einzelner Subgruppen auch einen Beitrag zur empirischen Überprüfung unterschiedlicher Erklärungsansätze für die geäußerte Absicht, eine City-Maut zu befürworten.

1.3.4. Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination

Der dritte Beitrag der Dissertation (Kapitel 4) schließt in methodischer Hinsicht an die vorigen Beiträge an, setzt thematisch jedoch nochmals einen anderen Fokus. So zielten die vorigen beiden Beiträge am Anwendungsbeispiel der Befürwortung einer möglichen City-Maut auf die inhaltlich motivierte Analyse von Heterogenität, sei es zur Modellierung eines vermuteten Effekts über den Wertebereich eines Moderators hinweg (Kapitel 2) oder der Identifikation von in unterschiedlichem Maße betroffenen Subgruppen (Kapitel 3). Im dritten Beitrag liegt der Fokus nun auf einer methodisch motivierten Analyse der Auswirkungen typischer Stichprobenverfahren auf die zu erwartende Heterogenität in der Zusammensetzung eines zufällig gezogenen Samples.

Am Beispiel eines Feldexperiments zu ethnischer Diskriminierung auf dem Wohnungsmarkt, bei dem Wohnungsanbieter:innen fiktive Besichtigungsanfragen zugesandt wurden, wird erforscht, welchen Einfluss verschiedene Aspekte der Stichprobenziehung auf die Zusammensetzung des erzielten Analysesamples haben. Zu denken ist dabei etwa an die üblicherweise vorgenommene Beschränkung auf nur ein Angebot pro Anbieter (Anbieter-Sampling) oder die Erhebung innerhalb eines relativ kurzen Zeitraums (Punkt-Sampling). Zu erwarten ist eine Überrepräsentation kleiner (privater) Anbieter und von Angeboten, die sich schon länger auf dem Markt befinden, mithin eine geringe Erfolgsquote aufweisen. In diesem Zusammenhang wird in der Literatur auch von einem Length Bias (van Es et al., 2000) gesprochen. Liegt infolge des gewählten Verfahrens ein verzerrtes Abbild des beobachteten Marktes vor, stellt sich die Frage, inwiefern das geschätzte Ausmaß von Diskriminierung oder auch die Identifikation zugrunde liegender Risikofaktoren von einer solchen Verzerrung beeinflusst sind. Die Idee ist also, die Kombination eines experimentellen Ansatzes mit weiteren Kontextinformationen (in diesem Fall zum zugrunde liegenden Markt) zur Kontrolle der Feldbedingungen zu nutzen, unter denen das Experiment durchgeführt wurde.

Die Verschiebung des Augenmerks auf die Untersuchung typischer Stichprobenziehungen bei Feldexperimenten zu Diskriminierung auf dem Wohnungsmarkt ergänzt die vorliegende Dissertationsschrift dabei in mehrerlei Hinsicht. Zum einen wird eine weitere Möglichkeit aufgezeigt, wie die Kombination von Experiment und Kontextinformationen in der empirischen Analyse produktiv genutzt werden kann.

Zum anderen wird der Blick für den Einsatz solcher Kombinationen in verschiedenen Anwendungsfeldern geweitet. Im Gegensatz zu den ersten beiden Beiträgen wird dabei auch nicht auf selbstberichtete Verhaltensabsichten abgezielt, sondern reales Verhalten beobachtet. Dabei könnten Fragen der ethnischen Diskriminierung am Wohnungsmarkt etwa vor dem Hintergrund künftig wohl zunehmender Flucht- und Migrationsbewegungen (unter anderem in Folge des Klimawandels; bspw. World Bank, 2016) an Relevanz gewinnen. Aber auch für sich betrachtet scheint die tiefgreifendere Erforschung von Diskriminierung aus sozialwissenschaftlicher Perspektive geboten, handelt es sich dabei doch um zentrale Fragen der Ungleichheitsforschung, die auf weitreichende Herausforderungen für das gesellschaftliche Zusammenleben verweisen. So deuten die Befunde aus bisherigen Studien einhellig auf das Vorliegen von Benachteiligungen gegenüber einer Minderheit zugehörigen Bewerber:innen. Das konkrete Ausmaß der Ungleichbehandlung in Form der berichteten Diskriminierungsrate (gemessen anhand seltenerer Antworten auf angefragte Besichtigungstermine) schwankt dabei allerdings sowohl zwischen untersuchten Ländern und Minderheiten als auch zwischen verschiedenen Studien deutlich (Auspurg et al., 2019a; Flage, 2018).

Der vorliegende Beitrag knüpft an dieser Stelle an und untersucht mögliche Ursachen, die zur beobachteten Variation der berichteten Diskriminierungsrate beitragen könnten. Dazu werden die Daten eines im Jahr 2015 auf dem deutschen Mietwohnungsmarkt durchgeführten Feldexperiments genutzt. In zwei jeweils einwöchigen Feldphasen im Mai und November wurden tägliche Zufallsstichproben von 500 angebotenen Zwei- bis Vier-Zimmer-Wohnungen von einer großen Online-Immobilienplattform gezogen. Dabei wurde nur jeweils maximal ein Angebot pro Anbieter berücksichtigt (Anbieter-Sampling). Für jedes der gezogenen Inserate wurden in randomisierter Reihenfolge Besichtigungsanfragen von zwei fiktiven Bewerbern versendet, wobei Absendername und verwendete E-Mailadresse jeweils einmal auf einen deutschen und einmal auf einen türkischen Bewerber deuteten (within-Design).[11] Die Analysen bleiben schlussendlich auf 2992 Angebote aus Westdeutschland beschränkt, für die $2992 \times 2 = 5984$ Anfragen versendet wurden.[12] Die wesent-

11 Weitere Variationen umfassten u.a. Informationen zu Erwerbsstatus und Einkommen, um auf Hinweise für statistische Diskriminierung testen zu können. Das durch den Namen signalisierte Geschlecht der Bewerber wurde nicht variiert. Zudem erfolgte der Versand der Anfragen mit inhaltlich leicht variierten Formulierungen und im Abstand von etwa einer Stunde, um das Risiko zu verringern, dass das Experiment entdeckt wird.

12 Die noch immer deutlich bestehenden Unterschiede zwischen Ost- und Westdeutschland sowohl hinsichtlich der Leerstandsquoten als auch in Bezug auf Einstellungen gegenüber Fremden

liche Stärke des Ansatzes besteht in der Kombination mit umfangreichen Informationen zur Marktsituation auf der Plattform, von der die Angebote gesampelt wurden. Hierzu wurden über den Zeitraum von etwa einem Jahr (und damit sowohl einige Wochen vor Beginn der ersten Feldphase im Mai als auch einige Wochen nach dem Ende der zweiten Feldphase im November), von wenigen kürzeren Unterbrechungen abgesehen, einmal täglich alle auf der Plattform inserierten Angebote erfasst (insgesamt 1.087.541 Angebote). In die Analysen gehen Informationen von 695.458 Angeboten mit unzensierter Laufzeit ein.

Empirisch zeigen sich die vermuteten Verzerrungen hinsichtlich der Zusammensetzung des Samples deutlich. So sind verglichen mit den im Markt zu erwartenden Anteilen kleinere (private) Anbieter im Sample deutlich überrepräsentiert. Ebenso sind zum Zeitpunkt des Experiments bereits länger inserierte Angebote überproportional im Sample vertreten. Diese systematischen Verzerrungen der Samplezusammensetzung sind zwar qua Design zu erwarten, werden allerdings dennoch üblicherweise nicht explizit diskutiert. Dabei stellt sich die Frage, inwiefern sie sich auch im beobachteten Ausmaß von Diskriminierung spiegeln. Insgesamt liegt die Netto-Diskriminierungsrate (die für den türkischen Bewerber geringere Antwortquote als für den deutschen Bewerber) bei rund 14%. Von kleineren Schwankungen abgesehen scheint diese vorgefundene Größenordnung weitgehend robust zu sein und liegt mit einem Range zwischen 12,5 und 16% (für die kleinsten und größten Anbieter, Ähnliches zeigt sich für die kürzesten und längsten Angebotsdauern) genau in dem Bereich auch in der Literatur berichteter Effektgrößen (Auspurg et al., 2019a; Flage, 2018). Gleichwohl erweisen sich die beobachteten Schwankungen der zu erwartenden Diskriminierungsraten als systematisch davon abhängig, welches Marktsegment betrachtet wird.[13] Womöglich könnten Verzerrungen der Samplezusammensetzung auch dazu beitragen, die sich aus dem Vergleich unterschiedlicher Studien ergebenden Länderunterschiede oder Veränderungen der Diskriminierungsrate im Zeitverlauf aufzuklären.

Um das Potential der Verknüpfung mit umfassenden Kontextinformationen des

adäquat zu berücksichtigen, würde vergleichsweise komplexe Regressionstechniken erfordern (Auspurg et al., 2019b), die hier nicht im Fokus der Analysen stehen.

13 Hinsichtlich möglicher Risikofaktoren für Diskriminierung finden sich Hinweise, dass vor allem für Angebote, die zum Zeitpunkt des Experiments erst seit Kurzem inseriert waren, die Effektstärken etwas größer sind und nur für diese Angebote finden sich Anzeichen für statistische Diskriminierung. Allerdings ist keiner dieser Moderationseffekte statistisch signifikant. Insgesamt scheint die Analyse von Risikofaktoren für Diskriminierung also nicht von den zu erwartenden Verzerrungen des Samples beeinträchtigt zu sein.

Marktes, aus dem die Angebote für das Feldexperiment gezogen wurden, noch weiter auszuschöpfen, wurden in einem abschließenden Schritt zusätzliche Simulationsanalysen durchgeführt. Diese erlauben es, den Einfluss verschiedener Stichprobenverfahren auf die zu erzielende Zusammensetzung des Samples basierend auf den echten Marktdaten zu untersuchen. Dabei wird entsprechend der theoretischen Erwartungen festgestellt, dass Anbieter-Sampling im Vergleich zu Sampling von Angeboten ungeachtet des Anbieters (Apartment-Sampling) durchweg zu einer klaren Überrepräsentation kleinerer (privater) Anbieter führt. Auch fallen für längere Erhebungszeiträume Verzerrungen der erfassten Angebotsdauern etwas geringer aus als bei sehr kurzen Feldphasen (Punkt-Sampling). Diese Überrepräsentation längerer Angebotsdauern verschwindet allerdings auch für sehr lange Erhebungsphasen nicht vollständig. Das liegt in der Verwendung eines prospektiven Sampling-Ansatzes begründet, bei dem über einen festgelegten Erhebungszeitraum hinweg iterativ mehrere Stichproben gezogen werden. Lediglich durch das retrospektive Ziehen eines gemeinsamen Samples aus dem gesamten Beobachtungszeitraum ließe sich eine solche Verzerrung der Angebotsdauern vermeiden. Das scheint allerdings in der Forschungspraxis kaum umsetzbar. Immerhin würde man wohl keine positive Rückmeldung auf ein bereits Wochen zurückliegendes Angebot erwarten. Aber auch der Suchprozess realer Bewerber:innen dürfte durch ein prospektives Verfahren besser wiedergespiegelt werden. Vor diesem Hintergrund liefern also auch die Erkenntnisse aus den simulierten Stichprobenziehungen einen wichtigen Beitrag dazu, die durch das gewählte Stichprobenverfahren bedingten Verzerrungen einordnen zu können und so Rückschlüsse auf Heterogenität in der realisierten Stichprobe ziehen zu können.

1.4. Fazit, Limitationen und Ausblick

Anliegen der vorliegenden Dissertation ist es, an verschiedenen empirischen Anwendungsbeispielen aufzuzeigen, wie der Einbezug von Kontextinformationen im sozialwissenschaftlichen Experiment dazu genutzt werden kann, methodischen Problemen bestehender Ansätze zu begegnen. Insbesondere stehen dabei verschiedene Möglichkeiten im Fokus, wie sich solche Kombinationen zunutze gemacht werden können, um Effektheterogenität adäquat zu berücksichtigen.

Das Vorgehen knüpft dabei in mehrerlei Hinsicht an die bestehende Literatur an. Zum einen lässt sich der Einzug des Experiments in die Soziologie zu weiten Teilen als eine Abfolge verschiedener Weiterentwicklungen und Kombination experimentel-

ler Designs mit zusätzlichen Datenquellen beschreiben (bspw. Gërxhani und Miller, 2022), die hier weitergeführt wird. Zum anderen bietet die gezielte Kombination experimenteller Ansätze mit Informationen zum Kontext, in dem ein Experiment durchgeführt wird, sowohl inhaltlich als auch methodisch aufschlussreiches Potential, wie die drei eigenen empirischen Anwendungsbeispiele illustrieren. Diese beinhalteten 1) die explizite Modellierung theoretisch zu erwartender Heterogenität in Form nicht-linearer Zusammenhänge bei der Befürwortung einer möglichen City-Maut über den Wertebereich eines Moderators hinweg (in diesem Fall des Umweltbewusstseins der Befragten), 2) die Verbesserung der Schätzung durch verringerte Heterogenität bei getrennter Analyse von Subgruppen (in unterschiedlichem Maße von einer möglichen City-Maut betroffener Befragter) und 3) die Identifikation von Heterogenität in der Zusammensetzung von Samples bei Feldexperimenten zu Diskriminierung auf dem Wohnungsmarkt (was die retrospektive Kontrolle von Feldbedingungen ermöglicht). Das Verfahren ist dabei insofern designbasiert, als jeweils vorab anhand theoretischer Vorüberlegungen eine Auswahl relevanter Kontextinformationen erfolgte, die schon in der Konzeption der Erhebung Berücksichtigung fand. Zu denken ist etwa an die beschriebene Marktbeobachtung bereits vor der eigentlichen Durchführung des Feldexperiments und darüber hinaus oder die Untersuchung von umwelt- und mobilitätsbezogenen Einstellungen und Verhaltensabsichten, die gezielt nicht nur im innerstädtischen Bereich, sondern auch unter Einbezug des Umlandes erfolgte. Ohne Berücksichtigung der Kontextinformationen wäre jeweils relevante Heterogenität des Treatmenteffekts übersehen worden.

Gleichwohl weisen die vorliegenden Beiträge auch Limitationen auf. So ist das vorgestellte Vorgehen der Kombination experimenteller Ansätze mit zusätzlichen Kontextinformationen beschränkt auf prinzipiell beobachtbare und auch tatsächlich beobachtete Kontextinformationen, die moderierend auf den jeweils interessierenden Zusammenhang wirken könnten. Damit können allerdings, und das mag die zentrale Einschränkung sein, mögliche Verzerrungen durch Endogenität qua Design nicht restlos ausgeschlossen werden. Konkret ist also beispielsweise im ersten Beitrag (Kapitel 2) sowohl das Umweltbewusstsein der Befragten als auch deren Einkommen lediglich beobachtet. Im Gegensatz zum Preis (der im faktoriellen Survey-Experiment in Form der zu erwartenden Mautgebühren variiert wurde) sind aus praktischen und forschungsethischen Erwägungen heraus weder Einstellungen noch das Einkommen der Befragten exogen variierbar. An dieser Stelle muss also theoriegeleitet argumentiert werden, dass keine anderen Faktoren den Zusammenhang beeinflussen und

entsprechend begründet werden inwiefern Heterogenität von Effekten zu erwarten ist. Die Überlegungen um die Low-Cost-Hypothese zielen dabei genau darauf ab, die variierende Stärke des Effekts des variierten Preises über den Wertebereich der moderierend wirkenden Einstellungen hinweg zu modellieren.

In ähnlicher Weise ist auch der zweite Beitrag (Kapitel 3) womöglich von eventuellen Endogenitätsproblemen betroffen. So wird aufgrund theoretischer Vorüberlegungen vermutet, dass der Wohnort der Befragten gekreuzt mit dem Autobesitz bei der Befürwortung einer möglichen City-Maut eine Rolle spielen könnte. Weder Wohnort noch Autobesitz konnten allerdings experimentell manipuliert werden.[14] Die zur Bewertung vorgelegten Beschreibungen fiktiver Mautmodelle wurden hingegen durchaus exogen variiert. Zumindest innerhalb der auch für sich betrachtet aufschlussreichen Subgruppen sollte also eine unverzerrte Schätzung des Treatmenteffekts möglich sein, sofern keine weiteren moderierend wirkenden Faktoren übersehen wurden. Die Schätzung eines insgesamt zu erwartenden Treatmenteffekts hängt schlussendlich von der korrekten Kenntnis der tatsächlichen Besetzung der Subgruppen in der Inferenzpopulation ab und ist daher mit Zurückhaltung zu interpretieren (vgl. Kohler et al., 2019; Tipton, 2022).

Eine weitere Limitation der ersten beiden Beiträge (Kapitel 2 und 3), die für die hier diskutierten Überlegungen zur Verbesserung der Belastbarkeit von Befunden experimenteller Studien relevant erscheint, ist deren Beschränkung auf Verhaltensabsichten. Unverbindlich geäußerte Absichten könnten womöglich abweichen von denjenigen Entscheidungen, die zum Beispiel im Rahmen eines im Ergebnis verbindlichen Bürger:innenentscheids zu erwarten wären. Diesem Einwand sieht sich der dritte Beitrag (Kapitel 4) nicht ausgesetzt. Hier wird tatsächliches Verhalten von Wohnungsanbieter:innen in Reaktion auf Besichtigungsanfragen unterschiedlicher Bewerber unter realen Feldbedingungen beobachtet.

Zentrale Einschränkung des dritten Beitrags (Kapitel 4) ist vielmehr, dass die umfassende Kontrolle der Feldbedingungen lediglich zur nachträglichen Einordnung der erzielten Befunde vor dem Hintergrund der bestehenden Literatur befähigt. So lässt sich im Nachhinein beurteilen, wie stark Verzerrungen des gezogenen Samples die Schätzung des interessierenden Treatmenteffekts beeinflusst haben. Wünschenswert wäre hingegen, bereits vor oder iterativ im Laufe der Erhebung entsprechende Korrekturen vornehmen zu können. Zu denken ist etwa an Gewichtungsfaktoren beim Sampling verschiedener Angebote, um unterschiedliche Auswahlwahrscheinlichkei-

14 Auch hier sind praktische und forschungsethische Erwägungen zu berücksichtigen.

ten (etwa nach Anbietergröße oder Angebotsdauer) auszugleichen.

An dieser Stelle ist eine prinzipielle Beschränkung der vorgeschlagenen Kombinationsmöglichkeiten experimenteller Verfahren mit Kontextinformationen zu betonen. So wurde in der vorliegenden Dissertationsschrift durchweg auf den theoriegeleiteten Einsatz solcher Kombinationen abgestellt. In allen drei Beiträgen wurde zunächst anhand theoretischer Vorüberlegungen abgewogen, welche Merkmale des jeweils relevanten Kontexts sich in welcher Weise auf einen interessierenden Zusammenhang auswirken könnten. Anschließend erfolgte deren Einbezug in Form von entsprechend angepassten Schätzungen des interessierenden Treatmenteffekts. Während ein solches Vorgehen insbesondere zum empirischen Test theoretischer Überlegungen geeignet erscheint, könnte es für explorative Analysen auch aufschlussreich sein, eventuelle Moderatoren aus den Daten selbst zu schätzen (bspw. mithilfe von Clusteranalysen; Tipton, 2013a). Konkret ergäbe sich eine sinnvolle Erweiterung, wenn neben den bereits berücksichtigten Einflüssen weitere bisher unbeachtete Faktoren ausschlaggebend für die im Experiment untersuchten Zusammenhänge wären. Bezogen auf das Anwendungsbeispiel der Befürwortung einer möglichen City-Maut ist beispielsweise an die Intensität der Autonutzung zu denken. Diese könnte sich jenseits der dichotomen Unterscheidung des Autobesitzes auf die Haltung gegenüber einer City-Maut auswirken. Hier explorativ entsprechende Muster finden zu können, setzt allerdings auch eine entsprechende Reichhaltigkeit der verfügbaren Datengrundlage voraus (die zwar nicht immer, aber wohl zunehmend gegeben sein dürfte; Couper, 2017).

Das verweist auf Ansatzpunkte für die künftige Forschung, die etwa mit noch feingliedrigeren Kontextinformationen das vorgeschlagene Vorgehen weiter vertiefen könnte. So beziehen beispielsweise die in Kapitel 3 vorgelegten Analysen den räumlichen Kontext lediglich in Form aggregierter Merkmale der Wohnumgebung ein. Detailliertere Informationen zur genauen Lage der Wohnung (etwa, ob sich diese direkt an einer vielbefahrenen Straße befindet) oder beispielsweise der zusätzliche Einbezug des Arbeitsortes sowie der Wegstrecke dorthin (auch mit öffentlichen Verkehrsmitteln als Ausweichoption zu den durch eine mögliche City-Maut anfallenden Gebühren) wären womöglich aufschlussreiche Ergänzungen. Dabei sind jeweils theoretische Vorüberlegungen zur Auswahl relevanter Faktoren erforderlich. Gerade an dieser Stelle scheinen sowohl qualitative als auch interdisziplinäre Ansätze vielversprechende Anhaltspunkte und Möglichkeiten zu bieten (bspw. Bruderer Enzler, 2017; Diekmann und Meyer, 2010; Grisolía et al., 2015; Rau und Scheiner, 2020). Immerhin verbessert sich die Präzision der Schätzung nur dann in bedeutendem

Maße, wenn die Subgruppen anhand geeigneter Kriterien gebildet werden. Das trifft auch für den Einbezug räumlicher Umgebungsmerkmale zu, wie beispielsweise Tipton (2022) betont. Neben erhöhter Präzision bei der Schätzung von Treatmenteffekten ergeben sich aus der Verknüpfung experimenteller Ansätze mit Informationen zum räumlichen Kontext auch Möglichkeiten zur Erforschung inhaltlich motivierter Fragestellungen, die etwa explizit auf regionale Heterogenität der Effekte von Umgebungsmerkmalen abzielen (für einen Überblick siehe beispielsweise Glenk et al., 2020; Sagebiel et al., 2020).

Die aus der vorliegenden Dissertation ableitbaren Desiderata für die künftige Forschung lassen sich im Kern in drei Punkten zusammenfassen. Erstens unterstreichen die vorgelegten Befunde die Forderung nach einer systematischeren Befassung mit Effektheterogenität auch im sozialwissenschaftlichen Experiment (vgl. Bryan et al., 2021). An verschiedenen empirischen Anwendungsbeispielen wurde gezeigt, dass trotz der gemeinhin als hoch eingestuften internen Validität experimenteller Ansätze die externe Validität und damit auch der Erkenntnisgewinn hinsichtlich zu testender theoretischer Konzepte zum Teil eingeschränkt sein mag, sollte hier substanzielle Variation übersehen werden. Dabei kann der gezielte Einbezug von Kontextinformationen dazu beitragen, relevante Faktoren zu berücksichtigen.

Zweitens verweisen die vorgelegten Befunde auf die Bedeutung der expliziten und detaillierten Benennung des Designs empirischer Analysen sowie der jeweils verwendeten Datengrundlage. Womöglich verleiten auch hier gerade experimentelle Ansätze aufgrund ihrer hohen internen Validität zu sehr weitreichenden Interpretationen vorgefundener Zusammenhänge. Inwiefern Rückschlüsse auf eine breitere Inferenzpopulation allerdings gerechtfertigt erscheinen, lässt sich nur vor dem Hintergrund weiterer Informationen angemessen beurteilen. Dabei ist etwa die Frage zentral, ob das jeweilige Sample eine zugrundeliegende Inferenzpopulation repräsentiert und falls nicht, welche theoretischen Annahmen über die Gültigkeit des vorgefundenen Zusammenhangs auch in der restlichen Population erforderlich sind.[15]

Drittens ist vor dem Hintergrund einer sich auch auf Teile sozialwissenschaftlicher Disziplinen erstreckenden Replikationskrise der Blick auf Anforderungen für künftige Replikationsbestrebungen zu richten (vgl. Auspurg und Brüderl, 2022; Otte et al.,

15 Letztlich weist auch das von Lundberg et al. (2021) vorgeschlagene Gerüst des sogenannten Estimands-Ansatzes in diese Richtung: Ohne klare Definition der zu schätzenden theoretischen Größe und der expliziten Begründung ihrer empirischen Entsprechung sowie der zur Identifikation erforderlichen Annahmen lässt sich nicht beurteilen, inwiefern vorgetragene Befunde belastbar sind.

2023). Dabei verweisen die vorliegenden Befunde darauf, dass jeweils zunächst zu prüfen ist, in welchem Kontext ursprüngliche Ergebnisse entstanden sind und welchen Geltungsanspruch sie erheben. Anschließend kann in zweierlei Hinsicht daran angeknüpft werden. Von einer direkten Replikation im engeren Sinn könnte entsprechend nur dann ausgegangen werden, wenn der im Original berichtete Effekt im selben Kontext geprüft wird. Lässt sich der interessierende Effekt auf diesem Wege nicht replizieren, ist dessen Belastbarkeit zu hinterfragen. Liegt hingegen eine Erweiterung auf einen breiteren Kontext, veränderte Rahmenbedingungen oder auch andere Anwendungsfälle vor, ist zunächst zu diskutieren, inwiefern tatsächlich mit dem zu replizierenden Effekt zu rechnen ist. Beispiele für initial starke Effekte, die in einem breiteren Kontext nicht repliziert werden konnten, finden sich etwa in der Literatur um die Wirksamkeit grüner Nudges (bspw. Allcott, 2015; Szaszi et al., 2022). Die nicht replizierbaren Effekte deuten dabei wohl zu weiten Teilen auf vormals unberücksichtigte Einflüsse des Kontexts, in dem sie ursprünglich beobachtet wurden. Anstatt eine Glaubwürdigkeitskrise zu befeuern, könnten sie also auch als Ausgangspunkt tiefgreifenderer Auseinandersetzungen mit den ihnen womöglich zugrunde liegenden Moderatoren dienen und so zur Weiterentwicklung von Theorien beitragen, wie etwa Bryan et al. (2021) argumentieren.

Die in der vorliegenden Dissertation angeführten empirischen Anwendungsbeispiele unterstreichen die zentrale Bedeutung theoretischer Annahmen über die Kontextabhängigkeit von Treatmenteffekten auch in diesem Sinne. Die vorgeschlagenen Varianten des Einbezugs von Kontextfaktoren sollen illustrieren, wie sich Überlegungen zur Heterogenität von Treatmenteffekten im sozialwissenschaftlichen Experiment empirisch testen lassen.

Literaturverzeichnis

Al-Ubaydli, Omar und John A. List. 2015. „Do Natural Field Experiments Afford Researchers More or Less Control Than Laboratory Experiments?“ *American Economic Review* 105:462–66.

Allcott, Hunt. 2011. „Social norms and energy conservation.“ *Journal of Public Economics* 95:1082–1095.

Allcott, Hunt. 2015. „Site selection bias in program evaluation.“ *The Quarterly Journal of Economics* 130:1117–1165.

Antusch, Samantha. 2022. „Climate change and human behaviour. Editorial.“ *Nature Human Behaviour* 6:1441–1442.

Antusch, Samantha und Lingxiao Yan. 2022. „Climate change and human behaviour.“ `https://www.nature.com/collections/icdbhbbibg` (Stand: 16.11.2022).

Auspurg, Katrin und Josef Brüderl. 2021. „Has the credibility of the social sciences been credibly destroyed? Reanalyzing the “Many Analysts, One Data Set” Project.“ *Socius* 7:23780231211024421.

Auspurg, Katrin und Josef Brüderl. 2022. „How to increase reproducibility and credibility of sociological research“, S. 512–527. In: Gërxhani, Klarita, Nan De Graaf und Werner Raub (Hrsg.) *Handbook of Sociological Science*. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing.

Auspurg, Katrin, Josef Brüderl, und Thomas Wöhler. 2019a. „Does Immigration Reduce the Support for Welfare Spending? A Cautionary Tale on Spatial Panel Data Analysis.“ *American Sociological Review* 84:754–763.

Auspurg, Katrin und Thomas Hinz. 2015a. *Factorial Survey Experiments*. Thousand Oaks, California: SAGE Publications.

Auspurg, Katrin und Thomas Hinz. 2015b. „Multifactorial experiments in surveys“, S. 294–320. In: Keuschnigg, Marc und Tobias Wolbring (Hrsg.) *Experimente in den Sozialwissenschaften: Soziale Welt - Sonderband 22*. Baden-Baden: Nomos.

Auspurg, Katrin und Ulf Liebe. 2011. „Choice-Experimente und die Messung von Handlungsentscheidungen in der Soziologie.“ *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 63:301–314.

Auspurg, Katrin, Andreas Schneck, und Thomas Hinz. 2019b. „Closed Doors Everywhere? A Meta-Analysis of Field Experiments on Ethnic Discrimination in Rental Housing Markets.“ *Journal of Ethnic and Migration Studies* 45:95–114.

Auspurg, Katrin, Andreas Schneck, und Fabian Thiel. 2020. „Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination.“ *Research in Social Stratification and Mobility* 65:100444.

Bader, Felix, Bastian Baumeister, Roger Berger, und Marc Keuschnigg. 2019. „On the Transportability of Laboratory Results.“ *Sociological Methods & Research* 50:1452–1481.

Bakdash, Jonathan Z. und Laura R. Marusich. 2022. „Left-truncated effects and overestimated meta-analytic means.“ *Proceedings of the National Academy of Sciences* 119:e2203616119.

Berger, Roger und Tobias Wolbring. 2015. „Kontrafaktische Kausalität und eine Typologie sozialwissenschaftlicher Experimente“, S. 39–57. In: Keuschnigg, Marc und Tobias Wolbring (Hrsg.) *Experimente in den Sozialwissenschaften: Soziale Welt - Sonderband 22*. Baden-Baden: Nomos.

Berger, Sebastian, Andreas Kilchenmann, Oliver Lenz, Axel Ockenfels, Francisco Schlöder, und Annika M. Wyss. 2022. „Large but diminishing effects of climate action nudges under rising costs.“ *Nature Human Behaviour* 6:1381–1385.

Best, Henning und Clemens Kroneberg. 2012. „Die Low-Cost-Hypothese. Theoretische Grundlagen und empirische Implikationen.“ *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 64:535–561.

Blau, Peter Michael. 1977. *Inequality and heterogeneity: A primitive theory of social structure*. New York: Free Press.

Bohner, Gerd und Lena E. Schlüter. 2014. „A Room with a Viewpoint Revisited: Descriptive Norms and Hotel Guests' Towel Reuse Behavior." *PLOS ONE* 9:e104086.

Braun, Norman. 1993. *Socially embedded exchange*. Frankfurt am Main: Peter Lang.

Breen, Richard. 2022. „Causal inference with observational data", S. 272–286. In: Gërxhani, Klarita, Nan De Graaf und Werner Raub (Hrsg.) *Handbook of Sociological Science*. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing.

Breen, Richard, Seongsoo Choi, und Anders Holm. 2015. „Heterogeneous causal effects and sample selection bias." *Sociological Science* 2:351–369.

Bruderer Enzler, Heidi. 2017. „Air travel for private purposes. An analysis of airport access, income and environmental concern in Switzerland." *Journal of Transport Geography* 61:1–8.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, Dustin Tingley, und Chagai M. Weiss. 2022. „Abstraction and Detail in Experimental Design." *American Journal of Political Science* (Advance online Publication):1–16.

Bryan, Christopher J., Elizabeth Tipton, und David S. Yeager. 2021. „Behavioural science is unlikely to change the world without a heterogeneity revolution." *Nature Human Behaviour* 5:980–989.

Campbell, Donald T. 1957. „Factors relevant to the validity of experiments in social settings." *Psychological Bulletin* 54:297–312.

Campbell, Donald T. und Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Coleman, James S. 1972. „Systems of social exchange." *Journal of Mathematical Sociology* 2:145–163.

Cook, Thomas D. und Donald T. Campbell. 1979. *Quasi-Experimentation. Design & Analysis Issues for Field Settings*. Chicago, IL: Rand Mc-Nally.

Couper, Mick P. 2017. „New Developments in Survey Data Collection." *Annual Review of Sociology* 43:121–145.

Cox, David Roxbee. 1958. *Planning of experiments*. New York: John Wiley & Sons.

Davies, Randall S., David D. Williams, und Stephen Yanchar. 2008. „The Use of Randomisation in Educational Research and Evaluation: A Critical Analysis of Underlying Assumptions." *Evaluation & Research in Education* 21:303–317.

Degtiar, Irina und Sherri Rose. 2023. „A Review of Generalizability and Transportability." *Annual Review of Statistics and Its Application* 10 (Advance online Publication).

Diekmann, Andreas. 2011. *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Rowohlts Enzyklopädie. Reinbek: Rowohlt.

Diekmann, Andreas und Reto Meyer. 2010. „Demokratischer Smog? Eine empirische Untersuchung zum Zusammenhang zwischen Sozialschicht und Umweltbelastungen." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62:437–457.

Diekmann, Andreas und Peter Preisendörfer. 1998. „Umweltbewußtsein und Umweltverhalten in Low- und High-Cost-Situationen. Eine empirische Überprüfung der Low-Cost-Hypothese." *Zeitschrift für Soziologie* 27:438–453.

Diekmann, Andreas und Peter Preisendörfer. 2003. „Green and Greenback. The behavioral effects of environmental attitudes in low-cost and high-cost situations." *Rationality and Society* 15:441–472.

Eliasson, Jonas und Lina Jonsson. 2011. „The unexpected "yes": Explanatory factors behind the positive attitudes to congestion charges in Stockholm." *Transport Policy* 18:636–647.

Falk, Armin und James J. Heckman. 2009. „Lab experiments are a major source of knowledge in the social sciences." *Science* 326:535–538.

Findley, Michael G., Kyosuke Kikuta, und Michael Denly. 2021. „External Validity." *Annual Review of Political Science* 24:365–393.

Flage, Alexandre. 2018. „Ethnic and Gender Discrimination in the Rental Housing Market: Evidence from a Meta-Analysis of Correspondence Tests, 2006-2017." *Journal of Housing Economics* 41:251–273.

Gelman, Andrew. 2015. „The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective.“ *Journal of Management* 41:632–643.

Gërxhani, Klarita, Nan de Graaf, und Werner Raub (Hrsg.). 2022. *Handbook of Sociological Science: Contributions to Rigorous Sociology*. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing.

Gërxhani, Klarita und Luis Miller. 2022. „Experimental sociology“, S. 309–323. In: Gërxhani, Klarita, Nan De Graaf und Werner Raub (Hrsg.) *Handbook of Sociological Science*. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing.

Glenk, Klaus, Robert J. Johnston, Jürgen Meyerhoff, und Julian Sagebiel. 2020. „Spatial Dimensions of Stated Preference Valuation in Environmental and Resource Economics: Methods, Trends and Challenges.“ *Environmental and Resource Economics* 75:215–242.

Goldstein, Noah J., Robert B. Cialdini, und Vladas Griskevicius. 2008. „A room with a viewpoint: Using social norms to motivate environmental conservation in hotels.“ *Journal of Consumer Research* 35:472–482.

Grisolía, José M., Francisco López, und Juan de Dios Ortúzar. 2015. „Increasing the acceptability of a congestion charging scheme.“ *Transport Policy* 39:37–47.

Hausman, Daniel M. und Brynn Welch. 2010. „Debate: To nudge or not to nudge.“ *Journal of Political Philosophy* 18:123–136.

Holland, Paul W. 1986. „Statistics and Causal Inference.“ *Journal of the American Statistical Association* 81:945–960.

Huber, Robert A., Michael L. Wicki, und Thomas Bernauer. 2020. „Public support for environmental policy depends on beliefs concerning effectiveness, intrusiveness, and fairness.“ *Environmental Politics* 29:649–673.

Hudgens, Michael G. und M. Elizabeth Halloran. 2008. „Toward causal inference with interference.“ *Journal of the American Statistical Association* 103:832–842.

Imai, Kosuke, Gary King, und Elizabeth A. Stuart. 2008. „Misunderstandings between experimentalists and observationalists about causal inference.“ *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171:481–502.

IPCC. 2022. „Climate Change 2022: Mitigation of Climate Change.“ Report, Cambridge, UK and New York, NY, USA: Cambridge University Press.

Jackson, Michelle und D. R. Cox. 2013. „The Principles of Experimental Design and Their Application in Sociology.“ *Annual Review of Sociology* 39:27–49.

Jaensirisak, Sittha, Mark Wardman, und Anthony D. May. 2005. „Explaining variations in public acceptability of road pricing schemes.“ *Journal of Transport Economics and Policy* 39:127–154.

Jiménez-Buedo, María und Luis M. Miller. 2010. „Why a Trade-Off? The Relationship between the External and Internal Validity of Experiments.“ *Theoria* 69:301–321.

Keiding, Niels und Thomas A. Louis. 2018. „Web-based enrollment and other types of self-selection in surveys and studies: consequences for generalizability.“ *Annual Review of Statistics and Its Application* 5:25–47.

Keuschnigg, Marc und Fabian Kratz. 2018. „Thou Shalt Recycle: How Social Norms of Environmental Protection Narrow the Scope of the Low-Cost Hypothesis.“ *Environment and Behavior* 50:1059–1091.

Kohler, Ulrich, Frauke Kreuter, und Elizabeth A. Stuart. 2019. „Nonprobability Sampling and Causal Analysis.“ *Annual Review of Statistics and Its Application* 6:149–172.

Lesko, Catherine R., Ashley L. Buchanan, Daniel Westreich, Jessie K. Edwards, Michael G. Hudgens, und Stephen R. Cole. 2017. „Generalizing study results: a potential outcomes perspective.“ *Epidemiology* 28:553.

Liebe, Ulf und Jürgen Meyerhoff. 2021. „Mapping potentials and challenges of choice modelling for social science research.“ *Journal of Choice Modelling* 38:100270.

Lundberg, Ian, Rebecca Johnson, und Brandon M. Stewart. 2021. „What is your estimand? Defining the target quantity connects statistical evidence to theory.“ *American Sociological Review* 86:532–565.

Maier, Maximilian, František Bartoš, T. D. Stanley, David R. Shanks, Adam J. L. Harris, und Eric-Jan Wagenmakers. 2022. „No evidence for nudging after adjusting for publication bias.“ *Proceedings of the National Academy of Sciences* 119:e2200300119.

Manzo, Gianluca (Hrsg.). 2021. *Research Handbook on Analytical Sociology*. Cheltenham, UK: Edward Elgar Publishing.

Mertens, Stephanie, Mario Herberz, Ulf J. J. Hahnel, und Tobias Brosch. 2022a. „Correction for Mertens et al., The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains.“ *Proceedings of the National Academy of Sciences* 119:e2204059119.

Mertens, Stephanie, Mario Herberz, Ulf J. J. Hahnel, und Tobias Brosch. 2022b. „The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains.“ *Proceedings of the National Academy of Sciences* 119:e2107346118.

Mertens, Stephanie, Mario Herberz, Ulf J. J. Hahnel, und Tobias Brosch. 2022c. „Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research.“ *Proceedings of the National Academy of Sciences* 119:e2202928119.

Milenković, Marina, Draženko Glavić, und Milica Maričić. 2019. „Determining factors affecting congestion pricing acceptability.“ *Transport Policy* 82:58–74.

Morgan, Stephen L. und Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. New York, NY: Cambridge University Press.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

Neyman, Jerzy Splawa. 1935. „Statistical Problems in Agricultural Experimentation.“ *Journal of the Royal Statistical Society* 2:107–154.

Neyman, Jerzy Splawa. 1990. „On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in

Roczniki Nauk Rolniczych Tom X (1923) 1-51 (Annals of Agricultural Sciences)." *Statistical Science* 5:465–472.

Opp, Karl-Dieter. 1969. „Das Experiment in den Sozialwissenschaften: Einige Probleme und Vorschläge für seine effektivere Verwendung." *Zeitschrift für die gesamte Staatswissenschaft / Journal of Institutional and Theoretical Economics* 125:106–122.

Otte, Gunnar, Tim Sawert, Josef Brüderl, Stefanie Kley, Clemens Kroneberg, und Ingo Rohlfing. 2023. „Gütekriterien in der Soziologie. Eine analytisch-empirische Perspektive." *Zeitschrift für Soziologie* (Advance online Publication).

Pearl, Judea und Elias Bareinboim. 2014. „External Validity: From Do-Calculus to Transportability Across Populations." *Statistical Science* 29:579–595.

Preisendörfer, Peter. 2014. „Umweltgerechtigkeit: Von sozial-räumlicher Ungleichheit hin zu postulierter Ungerechtigkeit lokaler Umweltbelastungen." *Soziale Welt* 65:25–45.

Rau, Henrike und Joachim Scheiner. 2020. „Sustainable Mobility: Interdisciplinary Approaches." *Sustainability* 12:9995.

Rentziou, Aikaterini, Christina Milioti, Konstantina Gkritza, und Matthew G. Karlaftis. 2011. „Urban road pricing: Modeling public acceptance." *Journal of Urban Planning and Development* 137:56–64.

Roe, Brian E. und David R. Just. 2009. „Internal and External Validity in Economic Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data." *American Journal of Agricultural Economics* 91:1266–1271.

Rohrschneider, Robert. 1988. „Citizens' attitudes toward environmental issues: Selfish or selfless?" *Comparative Political Studies* 21:347–367.

Ross, Ronald. 1916. „An application of the theory of probabilities to the study of a priori pathometry. Part I." *Proceedings of the Royal Society of London. Series A* 92:204–230.

Rubin, Donald B. 1974. „Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66:688–701.

Rubin, Donald B. 1980. „Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75:591–593.

Rubin, Donald B. 1986. „Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81:961–962.

Rubin, Donald B. 2008. „For Objective Causal Inference Design Trumps Analysis." *The Annals of Applied Statistics* 2:808–840.

Sagebiel, Julian, Klaus Glenk, und Jürgen Meyerhoff. 2020. „Does the place of residence affect land use preferences? Evidence from a choice experiment in Germany." *Bio-based and Applied Economics* 9:283–304.

Schelling, Thomas C. 1971. „Dynamic models of segregation." *Journal of Mathematical Sociology* 1:143–186.

Simmel, Georg. 2016. „Der Raum und die räumlichen Ordnungen der Gesellschaft", S. 9–17. In: Eigmüller, M. und G. Vobruba (Hrsg.): *Grenzsoziologie: Die politische Strukturierung des Raumes*. Wiesbaden: Springer Fachmedien.

Small, Mario L. und Laura Adler. 2019. „The Role of Space in the Formation of Social Ties." *Annual Review of Sociology* 45:111–132.

Sorokin, Pitirim A. 1928. „Arbeitsleistung und Entlohnung (Experimentelle Untersuchungen bei Kindern im Alter von 3–4 Jahren und von 13–14 Jahren)." *Kölner Vierteljahreshefte für Soziologie* 7:186–198.

Szaszi, Barnabas, Anthony Higney, Aaron Charlton, Andrew Gelman, Ignazio Ziano, Balazs Aczel, Daniel G. Goldstein, David S. Yeager, und Elizabeth Tipton. 2022. „No reason to expect large and consistent effects of nudge interventions." *Proceedings of the National Academy of Sciences* 119:e2200732119.

Thaler, Richard H. und Cass R. Sunstein. 2021. *Nudge: The final edition*. New York, NY: Penguin Books.

Thiel, Fabian. 2020. „Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72:429–453.

Thiel, Fabian. 2021. „Support for city road tolls: a question of self-interest?“ In: Franzen, Axel und Sebastian Mader (Hrsg.): *Research Handbook on Environmental Sociology*. Cheltenham, UK: Edward Elgar Publishing.

Thye, Shane R. 2014. „Chapter 3 - Logical and Philosophical Foundations of Experimental Research in the Social Sciences“, S. 53–82. In: Webster, Murray und Jane Sell (Hrsg.): *Laboratory Experiments in the Social Sciences*. San Diego: Academic Press.

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, und Thorsten Staake. 2018. „Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation.“ *Management Science* 64:1458–1476.

Tiefenbeck, Verena, Anselma Wörner, Samuel Schöb, Elgar Fleisch, und Thorsten Staake. 2019. „Real-time feedback promotes energy conservation in the absence of volunteer selection bias and monetary incentives.“ *Nature Energy* 4:35–41.

Tipton, Elizabeth. 2013a. „Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts.“ *Journal of Educational and Behavioral Statistics* 38:239–266.

Tipton, Elizabeth. 2013b. „Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments.“ *Evaluation Review* 37:109–139.

Tipton, Elizabeth. 2022. „Sample Selection in Randomized Trials With Multiple Target Populations.“ *American Journal of Evaluation* 43:70–89.

Tipton, Elizabeth und Laura R. Peck. 2017. „A Design-Based Approach to Improve External Validity in Welfare Policy Evaluations.“ *Evaluation Review* 41:326–356.

Tipton, Elizabeth, David S. Yeager, Ronaldo Iachan, und Barbara Schneider. 2019. „Designing Probability Samples to Study Treatment Effect Heterogeneity“, S. 435–456. In: Lavrakas, Paul J., Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw und Brady T. West (Hrsg): *Experimental Methods in Survey Research*. Hoboken, NJ: John Wiley & Sons.

Treischl, Edgar und Tobias Wolbring. 2021. „The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences.“ *Methods Data Analyses* 16:141–170.

Tutić, Andreas, Thomas Voss, und Ulf Liebe. 2017. „Low-Cost-Hypothese und Rationalität. Eine neue theoretische Herleitung und einige Implikationen.“ *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69:651–672.

van Es, Bert, Chris A. J. Klaassen, und Karin Oudshoorn. 2000. „Survival Analysis Under Cross-Sectional Sampling: Length Bias and Multiplicative Censoring.“ *Journal of Statistical Planning and Inference* 91:295–312.

VanderWeele, Tyler J. und Miguel A. Hernan. 2013. „Causal inference under multiple versions of treatment.“ *Journal of Causal Inference* 1:1–20.

VanderWeele, Tyler J., Alex R. Luedtke, Mark J. van der Laan, und Ronald C. Kessler. 2019. „Selecting Optimal Subgroups for Treatment Using Many Covariates.“ *Epidemiology* 30:334–341.

VanderWeele, Tyler J., Eric J. Tchetgen Tchetgen, und M. Elizabeth Halloran. 2014. „Interference and sensitivity analysis.“ *Statistical Science* 29:687.

Veltri, Giuseppe A. 2021. „Chapter 21: Experiments“. In: Manzo, Gianluca (Hrsg.): *Research Handbook on Analytical Sociology*. Cheltenham, UK: Edward Elgar Publishing.

Wiertz, Dingeman und Nan Dirk de Graaf. 2022. „The climate crisis: what sociology can contribute“, S. 475–492. In: Gërxhani, Klarita, Nan de Graaf und Werner Raub (Hrsg.): *Handbook of Sociological Science*. Cheltenham, UK und Northampton, MA, USA: Edward Elgar Publishing.

Wolbring, Tobias und Marc Keuschnigg. 2015. „Feldexperimente in den Sozialwissenschaften“, S. 222–250. In: Keuschnigg, Marc und Tobias Wolbring (Hrsg.): *Experimente in den Sozialwissenschaften: Soziale Welt - Sonderband 22*. Baden-Baden: Nomos.

World Bank. 2016. *High and Dry: Climate Change, Water, and the Economy*. Washington, DC: World Bank.

Xie, Yu, Jennie E. Brand, und Ben Jann. 2012. „Estimating Heterogeneous Treatment Effects with Observational Data.“ *Sociological Methodology* 42:314–347.

Yan, Lingxiao. 2022. „Behaviour as leverage. Editorial.“ *Nature Climate Change* 12:1069–1069.

Zhou, Xiang und Yu Xie. 2020. „Heterogeneous Treatment Effects in the Presence of Self-Selection: A Propensity Score Perspective.“ *Sociological Methodology* 50:350–385.

2. Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut

Dieses Kapitel ist erschienen als:

Thiel, Fabian. 2020. „Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72:429–453. DOI: https://doi.org/10.1007/s11577-020-00712-0

Zusammenfassung

Die Low-Cost-Hypothese (LCH) postuliert, dass der Effekt einer Einstellung auf einstellungskonformes Handeln mit steigenden Kosten sinkt. Tutić et al. (2017) formalisieren die theoretischen Implikationen der LCH mithilfe eines mikroökonomischen Modells. Sie veranschaulichen, dass sich die LCH anhand einer verbesserten Teststrategie bewährt, während sie basierend auf der vormals üblichen Modellierung mittels der expliziten Aufnahme eines Interaktionsterms scheitert. Die von ihnen präsentierten Beispiele erlauben jedoch nur einen eingeschränkten Test der LCH im weiteren Sinn eines Einkommenseffekts – nicht aber im engeren Sinn eines tatsächlichen Preiseffekts einstellungskonformen Handelns. Dieser Beitrag liefert eine wichtige Ergänzung, indem am Beispiel der Befürwortung einer City-Maut eine erste Prüfung der verbesserten Teststrategie der LCH im engeren Sinn vorgelegt wird. Datengrundlage ist ein faktorieller Survey (FS), in dem im Rahmen einer Bevölkerungsbefragung in München und vier Umlandgemeinden im Frühsommer 2018 mehr als 1.300 Personen über 5.300 fiktive Mautmodelle bewerteten. Kernbefund der vorliegenden Untersuchung ist, dass sich die LCH anhand der verbesserten Teststrategie für den betrachteten Anwendungsfall der Befürwortung einer City-Maut bewährt.

Abstract

The Low-Cost-Hypothesis (LCH) postulates that the effect of attitudes on attitudes conforming behavior decreases with rising costs of attitude conformity. Using a microeconomic model, Tutić et al. (2017) formalize the theoretical implications of the LCH. They illustrate that the LCH can fail using the previously common approach involving the explicit inclusion of an interaction term, while it holds using the enhanced test strategy. However, the presented examples allow only a limited test of the LCH in the more general sense of an income effect—yet not in the narrower sense of an actual price effect of attitudes conforming behavior. This article provides an important supplement by presenting the first test of the improved test strategy of the LCH in the narrower sense using the support for a city toll as an example. The data is based on a factorial survey (FS) in which more than 1,300 respondents rated more than 5,300 fictitious toll models as part of a population survey in Munich and four surrounding municipalities in the early summer of 2018. According to the presented analysis of the support for a city toll, the LCH holds using the improved test strategy.

2.1. Einleitung

Die Belastung mit Luftschadstoffen hat weitreichende gesundheitliche Folgen, die von verschiedenen Erkrankungen der Atemwege und des Herz-Kreislaufsystems bis hin zu vorzeitigen Todesfällen reichen (bspw. Apte et al., 2018; Lelieveld et al., 2019; Pope und Dockery, 2006). In den Jahren 2000 bis 2015 ist die Konzentration für Luftschadstoffe wie Stickstoffdioxid (NO_2) und Feinstaub ($PM_{2,5}$)[1] in deutschen Ballungsräumen zwar deutlich gesunken, die von der Weltgesundheitsorganisation erarbeiteten Wirkungsschwellen von 20 $\mu g/m^3$ (NO_2) bzw. 10 $\mu g/m^3$ ($PM_{2,5}$) im Jahresmittel werden jedoch nach wie vor an einer Vielzahl der Messstationen insbesondere in verkehrsnaher Lage überschritten (Umweltbundesamt, 2017, S. 46f.). Gerade in innerstädtischen Lagen scheinen wesentliche Minderungen der Belastung ohne eine Reduktion des Straßenverkehrsaufkommens kaum erreichbar. Da die individuelle Verkehrsmittelwahl hauptsächlich von materiellen Faktoren wie der Verfügbarkeit eines Autos sowie der Kostenersparnis im Vergleich zur Nutzung anderer Verkehrsmittel abhängt, sind nennenswerte Verhaltensanpassungen ohne Veränderungen der (gesetzlichen) Rahmenbedingungen jedoch nicht zu erwarten (Preisendörfer, 2000). Zudem bleibt offen, inwiefern sich ein entsprechendes Problembewusstsein und umweltorientierte Einstellungen in der Befürwortung politischer Maßnahmen zur Veränderung eben jener Rahmenbedingungen niederschlagen. Das gilt insbesondere vor dem Hintergrund, dass sich eine Veränderung der gesetzlichen Rahmenbedingungen möglicherweise direkt auf die Kosten individuellen Mobilitätsverhaltens auswirkt. Zu denken sei dabei neben generellen Fahrverboten für bestimmte Fahrzeugtypen (bspw. „Dieselfahrverbote“) etwa an die Einführung einer City-Maut. Es stellt sich also die Frage, wie umweltbezogene Einstellungen in Anbetracht individuell anfallender Kosten auf die Absicht, solche Maßnahmen zu befürworten, wirken. Ist die beabsichtigte Befürwortung umso größer, je ausgeprägter umweltbezogene Einstellungen, aber auch je geringer die aus einer Befürwortung resultierenden Kosten sind?

2.1.1. Bisherige Forschung zur LCH

Die Low-Cost-Hypothese (LCH) stellt einen verbreiteten Ansatz zur Erklärung des hier im Fokus stehenden Zusammenspiels von Einstellung, Kosten und Verhalten

1 Als Feinstaub werden Staubpartikel mit einem aerodynamischen Durchmesser < 10 μm (PM_{10}) sowie noch feinere Partikel mit einem aerodynamischen Durchmesser $< 2{,}5$ μm ($PM_{2,5}$) bezeichnet.

dar. Sie besagt, dass mit steigenden Kosten einstellungskonformen Handelns der Effekt der Einstellung auf das Verhalten sinkt. Im Anschluss an grundlegende Arbeiten von Diekmann und Preisendörfer (1998, 2003) wurde die LCH in einer Vielzahl empirischer Arbeiten getestet und immer wieder auch theoretisch diskutiert (etwa Best und Kroneberg, 2012; Braun und Franzen, 1995; Diekmann, 1998; Keuschnigg und Kratz, 2018). Umweltsoziologische Anwendungen umfassen dabei bspw. die monetäre Bewertung von Biodiversität in Wäldern (Liebe, 2007), die Umstellung auf ökologische Landwirtschaft (Best, 2008) oder die Nutzung von Ökostrom (Neumann und Mehlkop, 2018). Eine ganze Reihe von Arbeiten befasst sich etwa mit der Beteiligung an Recycling (Best, 2009b,a; Best und Kneip, 2011; Derksen und Gartrell, 1993; Diekmann und Preisendörfer, 1998, 2003; Keuschnigg und Kratz, 2018; Schultz und Oskamp, 1996). Darüber hinaus wurde die LCH auch an anderen Anwendungsfällen, wie etwa der Spendenbereitschaft an Hilfsorganisationen (Mayerl, 2010), der Durchsetzung sozialer Normen (Rauhut und Krumpal, 2008) oder beruflicher Umzugsentscheidungen in Paarbeziehungen (Auspurg et al., 2014) getestet.

Die breite empirische Rezeption der LCH hat jedoch eine ebenso umfangreiche Bandbreite teils widersprüchlicher empirischer Befunde hervorgebracht. Dabei orientieren sich die Arbeiten durchaus an der von Diekmann und Preisendörfer (2003) vorgeschlagenen Teststrategie, die neben den Haupteffekten die Aufnahme eines multiplikativen Terms zwischen der Einstellung und den Kosten einstellungskonformen Verhaltens vorsieht. Besonders deutlich wird die unbefriedigende Befundlage an den Arbeiten zur Beteiligung an Recycling. So stützen manche der Arbeiten die LCH (Derksen und Gartrell, 1993; Diekmann und Preisendörfer, 1998, 2003), während andere Studien keine Unterstützung für die LCH finden (Best, 2009b,a; Best und Kneip, 2011; Schultz und Oskamp, 1996).

Umso mehr sind Bestrebungen hervorzuheben, die durch die neuerliche Auseinandersetzung mit der theoretischen Herleitung der LCH sowie ihrer Modellierung in der empirischen Anwendung möglicherweise dazu beitragen, die uneinheitliche Befundlage aufzuklären. Keuschnigg und Kratz (2018) argumentieren etwa, dass die widersprüchlichen Ergebnisse durch unterschiedliche normative Erwartungen bedingt seien, die wiederum den Zusammenhang zwischen Einstellung und einstellungskonformem Verhalten moderieren. So finden sie den von der LCH postulierten Zusammenhang zwar für Recycling von Plastik, nicht jedoch für Recycling von Glas. Das führen sie darauf zurück, dass aufgrund unterschiedlicher normativer Erwartungen das Ausmaß des Umweltbewusstseins für verschiedene Wertstoffe nicht in gleichem

Maße verhaltensrelevant ist. Während Recycling von Glas normativ gewissermaßen „selbstverständlich“ geworden ist (Einstellung und Kosten spielen kaum mehr eine Rolle), ist dies für Plastik nicht der Fall (die von der LCH postulierten Zusammenhänge bleiben für die Entscheidung über die Beteiligung an Recycling relevant). Andererseits mag dies möglicherweise auch an zum Teil falschen Schlussfolgerungen aufgrund bisher üblicher Spezifikationen der Modelle liegen, wie Best und Kroneberg (2012) für binäre sowie Tutić et al. (2017) für kontinuierliche Verhaltensvariablen argumentieren.

Die von Tutić et al. (2017) vorgelegte Herleitung der LCH aus einem mikroökonomischen Modell legt für den Fall kontinuierlicher Verhaltensvariablen überzeugend dar, dass von der Aufnahme eines multiplikativen Terms der Einstellung mit den Kosten einstellungskonformen Handelns abzusehen ist.[2] Die von ihnen vorgeschlagene Modellierung bildet das Zusammenspiel von Einstellung und Kosten vielmehr implizit ab und sieht daher keine explizite Aufnahme eines solchen Interaktionsterms vor. Gegenüber der herkömmlichen Teststrategie zeichnet sich diese verbesserte Teststrategie insbesondere durch zwei wesentliche Vorteile aus. Erstens ist sie aufgrund ihrer Herleitung aus der Mikroökonomik theoretisch fundiert und erlaubt zweitens aufgrund der starken Formalisierung präzise Vorhersagen (siehe Abschnitt 2.2).

Tutić et al. (2017) zeigen zudem, dass die LCH anhand der herkömmlichen Teststrategie scheitern kann, während sie sich anhand der von ihnen vorgeschlagenen verbesserten Teststrategie bewährt. Dazu illustrieren sie die verbesserte Teststrategie anhand zweier Beispiele, testen dabei jedoch lediglich die LCH im weiteren Sinn einer beschränkten Ressourcenausstattung. Die Kostspieligkeit einstellungskonformen Handelns wird in den Beispielen also anhand des Einkommens operationalisiert. Mithin wird getestet, dass je geringer das Einkommen (und damit auch der finanzielle Handlungsspielraum) ist, desto geringer sei der Effekt der Einstellung.[3] Die LCH im engeren Sinn sieht hingegen die explizite Modellierung der Kosten einstellungskonformen Handelns (so gesehen im Sinne eines Preises einstellungskonformen Handelns) jenseits des Einkommenseffekts vor. Wie die Autoren auch selbst betonen,

2 Best und Kroneberg (2012) kommen für den Fall binärer Verhaltensvariablen ebenfalls zu dem Schluss, dass zur Prüfung der LCH kein variablenspezifischer Interaktionsterm aus Einstellung und Kosten einstellungskonformen Verhaltens („Einstellung × Kosten“) in das Modell aufgenommen werden sollte.

3 Die Autoren sprechen in diesem Zusammenhang auch von einer „Low-Income-Hypothese“, um deutlich zu machen, dass die Kostspieligkeit einstellungskonformen Handelns hier über die Höhe des verfügbaren Einkommens, nicht aber über die für einstellungskonformes Handeln anfallenden Kosten, gemessen wird.

scheinen beide Varianten der LCH inhaltlich bedeutsam. Nach bestem Wissen existiert jedoch bisher noch keine Prüfung der LCH im engeren Sinn eines Preiseffekts anhand der verbesserten Modellierung.

2.1.2. Der Beitrag eines neuerlichen Tests

In der vorliegenden Arbeit wird die von Tutić et al. (2017) vorgeschlagene Modellierung auf den Anwendungsfall der Befürwortung einer City-Maut zur Verbesserung der Luftqualität in größeren Städten übertragen. Es wird der Frage nachgegangen, ob der Einfluss des Umweltbewusstseins auf die Absicht, eine City-Maut zu befürworten, mit steigenden Kosten der Maut abnimmt, wie es anhand der LCH zu vermuten wäre. Dazu werden Daten einer Bevölkerungsbefragung verwendet. Neben Fragen zu umweltrelevanten Einstellungen sowie zur individuellen Verkehrsmittelnutzung umfasste die Erhebung einen faktoriellen Survey zur Befürwortung einer City-Maut in München. Den Befragten wurden jeweils vier mögliche Mautmodelle in Form von Vignetten zur Bewertung vorgelegt, wobei relevante Eigenschaften (Dimensionen), wie etwa die Höhe der Mautgebühren oder der Grad der durch die Maut voraussichtlich erreichten Luftverbesserung, experimentell variiert wurden. Insgesamt liegen über 5.000 Vignettenurteile von mehr als 1.300 Befragten vor, auf deren Basis die im Kontext der LCH relevanten Zusammenhänge geprüft werden.

Damit wird in zweierlei Hinsicht an die bestehende Literatur zur LCH angeknüpft. Zum einen wird die LCH im engeren Sinn eines Preiseffekts einstellungskonformen Handelns einer ersten empirischen Prüfung anhand der verbesserten Teststrategie unterzogen. Zum zweiten ist der gewählte Anwendungsfall gerade aufgrund der Relevanz gesetzlicher Rahmenbedingungen für die Kosten, welche bei der Nutzung verschiedener Verkehrsmittel anfallen, aufschlussreich. So wird geprüft, inwiefern die Bereitschaft zur einstellungskonformen Gestaltung gesetzlicher Rahmenbedingungen auch dann besteht, wenn daraus im Rahmen der individuellen Verkehrsmittelwahl zusätzliche Kosten erwachsen. Dabei wird an die Folgerungen von Diekmann und Preisendörfer (2003, S. 468) zu den Auswirkungen veränderter gesetzlicher Rahmenbedingungen auf individuelle, einstellungsrelevante Verhaltensentscheidungen angeknüpft. Im Fokus stehen hier jedoch nicht die in Reaktion auf veränderte Rahmenbedingungen neu zu treffenden individuellen Verhaltensentscheidungen, sondern der vorgelagerte Schritt der Zustimmung zu derlei Gesetzesreformen selbst. Insofern stellt es auch keine Einschränkung dar, dass anstatt auf tatsächliches Mobilitätsver-

halten auf die Befürwortung einer politischen Maßnahme abgezielt wird, die sich – sollte sie eingeführt werden – dauerhaft in tatsächlichen Kosten für bestimmte Mobilitätsformen niederschlägt. Vielmehr muss beachtet werden, dass die hier untersuchte abhängige Variable der „Befürwortung“ einer City-Maut als eine *Verhaltensabsicht* zu verstehen ist. Es handelt sich also nicht um tatsächliche Wahlentscheidungen, sondern um die Absicht, die zu bewertenden Mautmodelle (etwa im Rahmen einer politisch verbindlichen Bürgerbefragung auf Kommunalebene) entsprechend zu befürworten.

2.2. Theorie

Eine nur geringe Korrelation zwischen Einstellung und Verhalten ist ein häufig zu beobachtendes Phänomen. Zu dessen Erklärung wird, aufbauend auf grundlegende Arbeiten von Diekmann und Preisendörfer (1998, 2003), oftmals die LCH herangezogen. Sie besagt zunächst allgemein, dass der Einfluss der Einstellung auf das Verhalten mit steigenden Kosten einstellungskonformen Verhaltens sinkt. Während sich daraus die Richtung des vermuteten Zusammenhangs ergibt, ist jedoch keine Aussage über die funktionale Form des Zusammenhangs (also darüber, in welchem Maße der Einfluss der Einstellung bei steigenden Kosten sinkt) enthalten.

2.2.1. Die mikroökonomische Modellierung der LCH

Tutić et al. (2017) integrieren diese zunächst allgemein formulierte Hypothese unter Rückgriff auf Grundlagen der Mikroökonomie (vgl. bspw. Braun und Gautschi, 2011; Mas-Collel et al., 1995; Varian, 1992) in ein Modell der Nachfragetheorie.[4] Dabei wird von Akteuren ausgegangen, die Präferenzen über mögliche Handlungsalternativen aufweisen. Die Auswahl zur Verfügung stehender Handlungsalternativen ist jedoch Restriktionen unterworfen. Akteure verhalten sich rational in dem Sinne, dass sie diejenige Handlungsalternative bzw. Kombination aus Handlungsalternativen wählen, die sie unter den gegebenen Restriktionen am höchsten bewerten (die ihnen mithin den größten Nutzen stiftet). Sofern die Präferenzen vollständig und

4 Die Autoren verweisen auch auf alternative Ansätze zur Formalisierung der LCH, wie sie von Diekmann (1998) und von Braun und Franzen (1995) vorgelegt wurden, die jedoch gewisse Schwächen aufweisen. An dieser Stelle wird lediglich das hier auch verwendete Modell von Tutić et al. (2017) skizziert. Für weitere Details sowie eine Diskussion der alternativen Ansätze siehe Tutić et al. (2017).

transitiv sind, lässt sich die Präferenzordnung als eine ordinale Nutzenfunktion verstehen. Die im Rahmen der LCH relevanten Zusammenhänge lassen sich sodann mithilfe einer Cobb-Douglas-Nutzenfunktion darstellen, wobei die Stärke der Einstellung („attitude“) α als Exponent des Ausmaßes einstellungskonformen Handelns einbezogen wird und so intrinsische Motive zu einstellungskonformem Handeln zu reflektieren vermag. Der individuelle Nutzen setzt sich schließlich aus dem (jeweils kontinuierlich gemessenen) Ausmaß einstellungskonformen ($x_a \geq 0$) und einstellungsirrelevanten Handelns ($x_{\neg a} \geq 0$) zusammen. Letzteres umfasst sämtliche Handlungsweisen, die nicht als einstellungskonform angesehen werden können. Dabei sind neben explizit im Widerspruch zur Einstellung stehenden Verhaltensweisen auch solche gemeint, die gewissermaßen als „neutral“ zu betrachten sind. Es wird nunmehr unterstellt, dass sich Akteure verhalten als ob sie diese Nutzenfunktion maximieren würden. Unter Berücksichtigung, dass die Wahl der jeweiligen Handlungsalternative unter Restriktionen in Form eines begrenzten Einkommens $m > 0$ sowie der Preise $p_a > 0$ für einstellungskonformes und $p_{\neg a} > 0$ für einstellungsirrelevantes Handeln getroffen wird, ergibt sich die Nachfragefunktion nach einstellungskonformem Handeln als

$$x_a^*(\alpha, m, p_a) = \alpha \frac{m}{p_a}, \tag{2.1}$$

wobei x_a^* das optimale Ausmaß einstellungskonformen Handelns bezeichnet. Anhand der Nachfragefunktion lassen sich die im Rahmen der LCH relevanten Zusammenhänge zwischen Einstellung, Kosten und der gewählten Handlungsalternative in der vermuteten Weise beschreiben. So ergibt sich ein positiver Effekt der Einstellung, ein negativer Effekt der Kosten einstellungskonformen Handelns sowie die kostenabhängige Stärke des Einstellungseffekts.[5] Neben der theoretischen Fundierung der im Rahmen der LCH erwarteten Zusammenhänge erlaubt das Modell zudem präzise Prognosen, da der funktionale Zusammenhang zwischen den relevanten Variablen klar spezifiziert ist.

Durch Logarithmieren der Nachfragefunktion erhält man schließlich die Gleichung

$$\ln x_a^*(\alpha, m, p_a) = \ln \alpha + \ln m - \ln p_a, \tag{2.2}$$

mithilfe derer die postulierten Zusammenhänge in Abb. 2.1 graphisch veranschau-

5 Die vermuteten Zusammenhänge gehen aus entsprechenden partiellen Ableitungen hervor (siehe Übersicht in Tab. 2.A1).

licht werden. Das anhand des Modells zu erwartende Ausmaß einstellungskonformen Handelns ist jeweils schematisch über die Stärke der Einstellung (links), die Höhe des Preises einstellungskonformen Handelns (mittig) und die Höhe des verfügbaren Einkommens (rechts) abgetragen (durchgezogene Linie). Die gepunktete und die gestrichelte Linie stellen dar, wie sich der Zusammenhang verschiebt, wenn die jeweilige Einflussgröße unter Konstanthalten der anderen Variablen erhöht wird. So wird etwa deutlich, dass der Anstieg des Ausmaßes einstellungskonformen Handelns mit zunehmender Stärke der Einstellung abnimmt, was der Vorstellung eines abnehmenden Grenznutzens entspricht. Wird lediglich der Preis einstellungskonformen Handelns erhöht, verschiebt sich die Linie nach unten – das jedoch stärker bei stärker ausgeprägter Einstellung, da hier auf deutlich höherem Niveau einstellungskonformes Handeln „nachgefragt“ wird. Wird statt des Preises ausschließlich das verfügbare Einkommen erhöht, verschiebt sich die Kurve entsprechend nach oben.

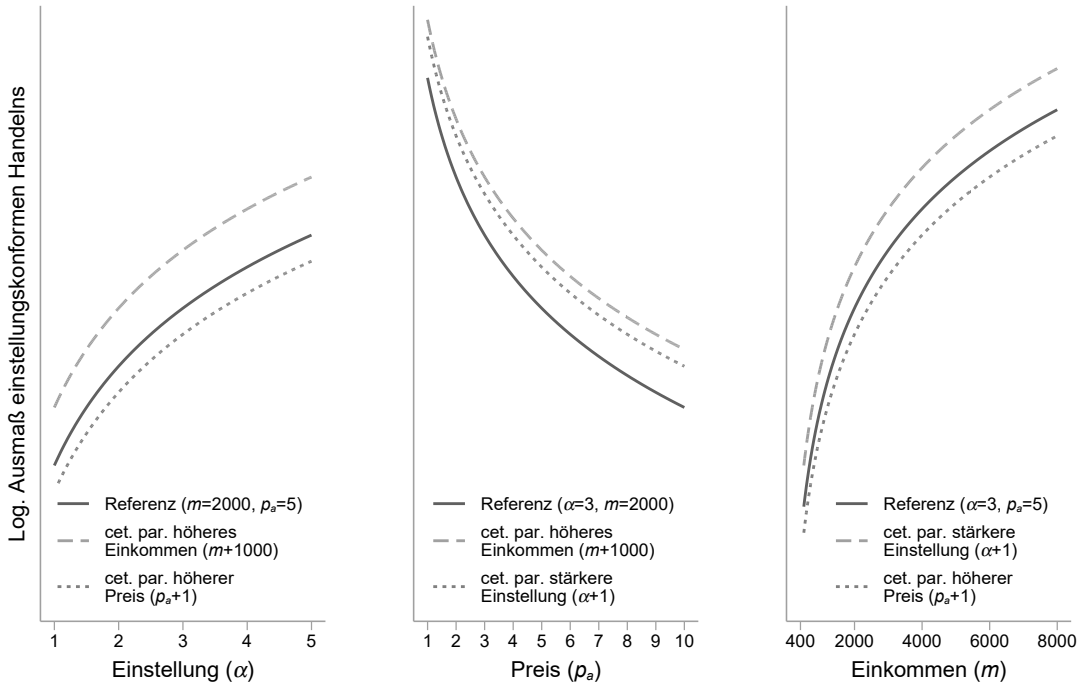

Abbildung 2.1.: Schematische Darstellung der anhand der Low-Cost-Hypothese zu erwartenden Zusammenhänge; dargestellt ist das vorhergesagte logarithmierte Ausmaß einstellungskonformen Verhaltens über die Einstellung (α; in Skalenpunkten; links), den Preis (p_a; in €; mittig) und das Einkommen (m; in €; rechts)

2.2.2. Bisherige Prüfung der neuen Modellierung

Tutić et al. (2017) illustrieren die verbesserte Teststrategie anhand einer Reanalyse der Datengrundlage zweier bereits publizierter Studien, einerseits zur monetären Bewertung der Biodiversität in Wäldern (Liebe, 2007) und andererseits zu Einstellungen gegenüber Tieren (Liebe und Jahnke, 2017). In beiden Fällen scheitert die LCH unter Verwendung der vormals üblichen Teststrategie, welche die Aufnahme eines expliziten Interaktionsterms zwischen der Einstellung (Umweltbewusstsein bzw. Bewusstsein für Tierwohl) und der Kostspieligkeit einstellungskonformen Verhaltens (geringeres Einkommen) fordert. Anhand der neuen (theoretisch fundierten) Teststrategie finden Tutić et al. (2017) hingegen durchaus Unterstützung für die LCH.

Dabei muss jedoch eingewandt werden, dass die Reanalyse bereits publizierten Materials gerade zum Test einer verbesserten Teststrategie und vor dem Hintergrund der teils widersprüchlichen Befundlage (siehe Abschnitt 2.1.1) zwar sinnvoll scheint, damit aber auch Beschränkungen der Datengrundlage in Kauf genommen werden müssen. So wird in den genannten Anwendungsfällen die Zahlungsbereitschaft für Biodiversität im Wald bzw. Spenden an eine Tierschutzorganisation als Ausmaß einstellungskonformen Handelns betrachtet. Die Kosten einstellungskonformen Handelns werden sodann über das Haushaltsäquivalenzeinkommen operationalisiert. Dies stellt in zweierlei Hinsicht ein Problem dar. Zum einen ist die lediglich über das Haushaltsäquivalenzeinkommen erfasste Kostspieligkeit einstellungskonformen Handelns (ohne eine exogen variierte Kostenkomponente in Form eines Preises einstellungskonformen Handelns) endogen, was potentiell zu verzerrten Ergebnissen führen kann. Zum anderen ermöglicht dieses Vorgehen lediglich einen Test der LCH im weiteren Sinne eines beschränkten Einkommens (die Autoren sprechen in diesem Zusammenhang selbst auch von einer Low-Income-Hypothese (LIH)), nicht aber der LCH im engeren Sinne eines Preises einstellungskonformen Handelns. Nach bestem Wissen existiert eine solche Prüfung der LCH (im engeren Sinn) anhand der von Tutić et al. (2017) vorgeschlagenen Teststrategie bisher nicht. Die vorliegende Studie soll am Anwendungsfall der Bereitschaft, eine City-Maut zur Verbesserung der Luftqualität zu befürworten, einen ersten Test der LCH anhand der verbesserten Teststrategie liefern.

2.2.3. Das Anwendungsbeispiel einer City-Maut

Das Modell lässt sich unter anderem auf den vorliegenden Anwendungsfall der Befürwortung einer City-Maut zur Verbesserung der Luftqualität übertragen. So kann etwa vermutet werden, dass mit einem hohen Umweltbewusstsein auch eine erhöhte Bereitschaft zu umweltbewussten Verhaltensweisen einhergeht. Dabei wird angenommen, dass der intrinsische Nutzen aus umweltbewussten Verhaltensweisen umso größer ist, je ausgeprägter das Umweltbewusstsein ist. Im Fall von Umweltschäden, wie sie unter anderem durch die straßenverkehrsbedingten Emissionen verursacht werden, müsste sich bei ausgeprägtem Umweltbewusstsein ein hoher intrinsischer Nutzen aus Verhaltensweisen, die zur Beseitigung der Umweltschäden beitragen, ergeben. Davon ausgehend, dass eine City-Maut zu einer Verringerung der Emissionen und damit zur Verbesserung der (innerstädtischen) Luftqualität beiträgt, wird also ein positiver Effekt des Umweltbewusstseins auf die Absicht, eine City-Maut zu befürworten, erwartet (vgl. Abb. 2.1, links).

Da Akteure allerdings nur über begrenzte Ressourcen verfügen, kann eine mit steigender Kostspieligkeit abnehmende Bereitschaft zur Befürwortung einer City-Maut erwartet werden. Auch Personen mit ausgeprägtem Umweltbewusstsein verfügen nur über begrenzte Ressourcen, können mithin ihre Präferenzen nicht unmittelbar umsetzen, sondern nur soweit es die Rahmenbedingungen zulassen. Somit kann schließlich ein mit steigender Kostspieligkeit abnehmender Einfluss des Umweltbewusstseins auf die Befürwortung der Maut erwartet werden. Dabei spielt einerseits die Höhe der Mautgebühren im Sinne eines Preiseffekts eine Rolle (vgl. Abb. 2.1, mittig), andererseits aber auch die Ressourcenausstattung der Akteure im Sinne eines Einkommenseffekts (vgl. Abb. 2.1, rechts). Zusammenfassend lässt sich festhalten: Je höher die Mautgebühren sind, aber auch je geringer das zur Verfügung stehende Einkommen ist, desto geringer ist der Einfluss des Umweltbewusstseins. Im vorliegenden Beitrag wird untersucht, ob sich diese Zusammenhänge in der erwarteten Weise zeigen (sie also den im theoretischen Modell formal dargestellten Beziehungen folgen; siehe auch Abschnitt 2.3.2).

Basierend auf diesen Ausführungen lässt sich zudem argumentieren, dass die vermuteten Zusammenhänge insbesondere für die Subgruppe an Personen, die in ihrem Mobilitätsverhalten direkt von einer möglichen Maut betroffen wären, bestehen sollten. So dürfte die Höhe der Mautgebühren etwa gerade für Personen, die (häufig) mit einem Kraftfahrzeug in einer der möglichen Mautzonen fahren, entscheidungsre-

levant sein. Diese hätten bei Einführung einer Maut individuelle Kosten zu tragen, während sich die Einführung einer Maut für andere Subgruppen (etwa Personen, die derzeit über kein selbstgenutztes Kraftfahrzeug verfügen) primär in veränderten Rahmenbedingungen der künftigen Verkehrsmittelwahl niederschlüge. In den empirischen Analysen wird also weiterhin geprüft, ob die Ergebnisse für unterschiedliche Messungen der Kostenkomponente substanziell robust bleiben. Abschließend werden die mithilfe der verbesserten Modellierung gewonnenen Ergebnisse solchen gegenübergestellt, die auf der herkömmlichen Teststrategie basieren.

2.3. Daten und Methoden

Die nachfolgenden Analysen basieren auf Daten einer im Frühsommer 2018 postalisch durchgeführten Bevölkerungsbefragung. Die Befragung wurde begleitend zur Methodenausbildung am Institut für Soziologie der Ludwig-Maximilians-Universität München durchgeführt. Hierzu wurde aus den jeweiligen Melderegistern eine Zufallsstichprobe der Wohnbevölkerung Münchens sowie einiger Umlandgemeinden mit hoher Pendelverflechtung zu München gezogen. Für München wurden 3.400 Personen gezogen, für die Umlandgemeinden jeweils 500 aus Gröbenzell, Landshut, Poing und Rosenheim. Insgesamt wurden 1.335 der 5.400 verschickten Fragebögen ausgefüllt zurückgesandt. Abzüglich 162 unzustellbarer Fragebögen wird somit eine (bereinigte) Ausschöpfungsquote von 26% realisiert. Aufgrund teilweise fehlender Angaben beschränkt sich das Analysesample auf 4.317 Vignettenurteile von 1.102 Personen.

2.3.1. Operationalisierung

Im Rahmenfragebogen wurde Umweltbewusstsein anhand der von Diekmann und Preisendörfer (1998) entwickelten Skala zur Erfassung des allgemeinen Umweltbewusstseins erhoben. Damit wird die in der Literatur gängige Messung aufgegriffen und so die Vergleichbarkeit von Ergebnissen ermöglicht. Die Skala umfasst neun Items, die sich in eine affektive, eine kognitive und eine konative Komponente einteilen lassen. Die affektive Komponente bezieht sich auf emotionale Betroffenheit durch Umweltprobleme, die kognitive Komponente beschreibt die Einsicht, dass ein (durch Menschen verursachtes) Umweltproblem besteht und die konative Komponente zielt auf die grundsätzliche Bereitschaft zu individuellen oder kollektiven Handlungen zur Bekämpfung des Problems. Alle Items werden auf einer fünfstufigen Rating-Skala

erfasst. Basierend auf einer Faktorenanalyse ergibt sich unter Einbezug aller neun Items eine eindimensionale Messung des Umweltbewusstseins, bei der 38% der Varianz der Items erklärt werden und alle Faktorladungen über 0,4 liegen. Die neun Items werden also zu einem additiven Index zusammengefasst (Cronbachs $\alpha = 0,85$).[6] Dieser weist einen Mittelwert von 3,77 (mit einem Minimum von 1, einem Maximum von 5 und einer Standardabweichung von 0,75) auf.

Die Befürwortung einer möglichen City-Maut in München wurde mithilfe eines faktoriellen Surveys erhoben. Dieses Vorgehen erlaubt die Vorteile eines experimentellen Designs im Rahmen einer Befragung zu nutzen (siehe bspw. Auspurg und Hinz, 2015). Den Befragten wurden jeweils vier hypothetische Mautmodelle (Vignetten) zur Bewertung vorgelegt, deren Merkmale (Dimensionen) sich in der jeweiligen Ausgestaltung unterscheiden. Die variierten Dimensionen umfassen neben der Höhe der Mautgebühren u.a. das zu erwartende Ausmaß der Luftverbesserung oder den Geltungsbereich der Maut. Es wurde, den methodischen Empfehlungen zur Erstellung faktorieller Surveys folgend, eine effiziente Auswahl der verwendeten 1.200 unterschiedlichen Vignetten getroffen. Dabei werden die Vignetten so gewählt, dass bei möglichst geringen Korrelationen zwischen den Dimensionen maximale Varianz der Level vorliegt. So wird die Präzision der geschätzten Effekte der Vignettendimensionen auf das Vignettenurteil erhöht (bspw. Auspurg und Hinz, 2015).[7] Durch die randomisierte Zuweisung der Vignetten (und damit der experimentellen Stimuli) ist zudem sichergestellt, dass die Vignettendimensionen nicht mit Merkmalen der Befragten korrelieren.

Abb. 2.2 zeigt eine Beispielvignette sowie die elfstufige Rating-Skala, inwiefern das beschriebene Mautmodell eingeführt werden sollte (-5 = „auf keinen Fall“ bis 5 =

6 Die Items umfassen die folgenden Aussagen: (1) „Es beunruhigt mich, wenn ich daran denke, unter welchen Umweltverhältnissen unsere Kinder und Enkelkinder wahrscheinlich leben müssen.“ (2) „Wenn wir so weitermachen wie bisher, steuern wir auf eine Umweltkatastrophe zu.“ (3) „Wenn ich Zeitungsberichte über Umweltprobleme lese oder entsprechende Fernsehsendungen sehe, bin ich oft empört und wütend.“ (4) „Es gibt Grenzen des Wachstums, die unsere industrialisierte Welt schon überschritten hat oder sehr bald erreichen wird.“ (5) „Derzeit ist es immer noch so, dass sich der größte Teil der Bevölkerung wenig umweltbewusst verhält.“ (6) „Nach meiner Einschätzung wird das Umweltproblem in seiner Bedeutung von vielen Umweltschützern stark übertrieben.“ (7) „Es ist immer noch so, dass die Politiker/innen viel zu wenig für den Umweltschutz tun.“ (8) „Zugunsten der Umwelt sollten wir alle bereit sein, unseren derzeitigen Lebensstandard zu senken.“ (9) „Umweltschutzmaßnahmen sollten auch dann durchgesetzt werden, wenn dadurch Arbeitsplätze verloren gehen.“

7 Technisch ausgedrückt wurde ein sogenanntes D-effizientes Design gewählt, bei dem auch alle Zweifach- und Dreifachinteraktionen zwischen den Dimensionen orthogonalisiert wurden. Für eine Übersicht der variierten Dimensionen siehe Tab. 2.A2.

„auf jeden Fall“). Um in der weiteren Analyse das Vignettenurteil in logarithmierter Form verwenden zu können, wird es auf einen Wertebereich von $1-11$ reskaliert, wobei hohe Werte weiterhin positivere Bewertungen widerspiegeln. Der Grad der zum jeweiligen Mautmodell geäußerten Zustimmung wird als einstellungskonforme Verhaltensabsicht, das Mautmodell auch im Rahmen einer (verbindlichen) Abstimmung zu befürworten, angesehen. Wie bereits angedeutet, handelt es sich dabei nicht um tatsächliches Verhalten. Mithin fallen auch die damit einhergehenden Kosten nicht unmittelbar an. Die Einführung einer City-Maut brächte allerdings durchaus umfassende Auswirkungen auf verhaltensrelevante Rahmenbedingungen mit sich. Es kann also argumentiert werden, dass die Absicht, ein solches Mautmodell zu befürworten, auch die Bereitschaft umfasst, tatsächlich anfallende (Verhaltens-)Kosten zu tragen und so als einstellungskonforme, mittelbar kostspielige Verhaltensabsicht betrachtet werden kann.

Dieses Vorgehen scheint auch empirisch gerechtfertigt. Zwar existieren bisher nur wenige und zudem teils widersprüchliche Befunde zur Validität der in Surveyexperimenten gemessenen Verhaltensabsichten in Bezug auf tatsächliches Verhalten. So finden u.a. Hainmueller et al. (2015) kongruente Ergebnisse für Verhaltensabsicht und tatsächliches Verhalten, während etwa Findley et al. (2017) deutliche Unterschiede zwischen den Messungen berichten. Insgesamt scheint sich jedoch abzuzeichnen, dass bei Verhaltensweisen, die kaum mit sozialer Erwünschtheit besetzt sind, von einer hohen Übereinstimmung zwischen Verhaltensabsicht und tatsächlichem Verhalten (zumindest hinsichtlich der Richtung und relativen Stärke von Effekten) ausgegangen werden kann (Petzold und Wolbring, 2019, S. 8). Im vorliegenden Fall kann zwar durchaus argumentiert werden, dass die Befürwortung geeigneter Maßnahmen zur (innerstädtischen) Luftverbesserung sozial erwünscht sei, nicht jedoch welche Maßnahme (etwa eine City-Maut oder Dieselfahrverbote) die zu befürwortende ist.

Inwieweit die einstellungskonforme Bereitschaft zur Befürwortung der Maut als eine kostspielige Verhaltensabsicht angesehen werden kann, richtet sich dabei sowohl nach dem zur Verfügung stehenden Einkommen, operationalisiert anhand des Haushaltsnettoäquivalenzeinkommens mit einem Mittelwert von 2.518,78 € (einem Minimum von 400, einem Maximum von 20.000 und einer Standardabweichung von 1.520,88), als auch nach der Höhe des Preises einstellungskonformen Handelns in Form der anfallenden Mautgebühren. Letztere wurden in Abstufungen zwischen 1 und 10 € pro Tag variiert.[8]

8 Die Anzahl möglicher Ausprägungen (Level) der Höhe der Mautgebühren wurde zwischen zwei

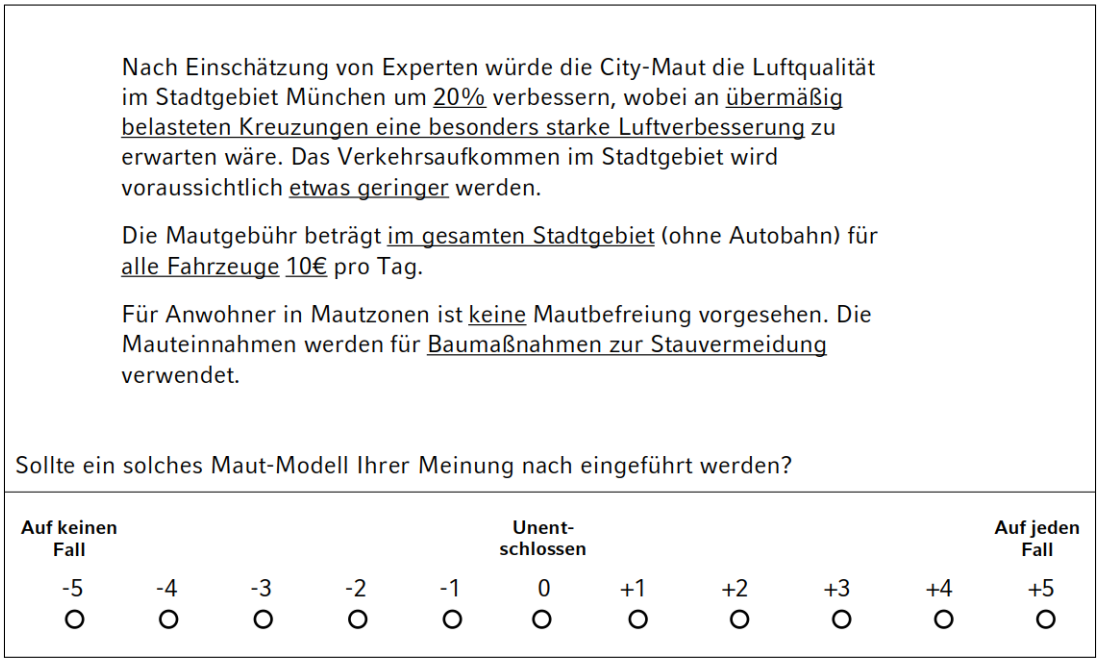

Abbildung 2.2.: Beispielvignette mit Antwortskala (variierte Dimensionen unterstrichen)

Im Rahmen weiterer Robustheitsanalyen wird zudem eine alternative Operationalisierung der Kosten herangezogen, die dem Umstand Rechnung trägt, dass die Kostspieligkeit einer City-Maut insbesondere für Personen, die in ihrem jeweiligen Mobilitätsverhalten direkt betroffen wären, entscheidungsrelevant sein sollte. Hierbei stehen also die bei gegebenem Mobilitätsverhalten tatsächlich anfallenden Mautgebühren im Fokus. Die Höhe der individuell zu erwartenden Kosten errechnet sich dabei aus der Höhe der Mautgebühren multipliziert mit der Häufigkeit, in der Befragte ein Kraftfahrzeug in einer der möglichen Mautzonen nutzen.[9] Dieser Kostenterm wird für Personen, für deren Kraftfahrzeug eine entsprechende Maut nicht gelten soll

zufällig ausgewählten Teilsplits von jeweils 50% der Befragten variiert. Für einen Teil der Befragten variierte die Höhe zwischen 1, 5 und 10 € pro Tag, für den anderen Teil der Befragten wurden zudem weitere Abstufungen von 3 € und 7 € pro Tag aufgenommen, um einen möglichen nichtlinearen Effekt der Mautgebühren besser schätzen zu können. Im Folgenden sind die gepoolten Ergebnisse dargestellt. Die berichteten Zusammenhänge zeigen sich substanziell aber auch für getrennte Analysen der beiden Teilsplits (mit viel bzw. wenig Variation der Höhe der Mautgebühren).

9 Die Häufigkeit der Autonutzung wird über die Angabe, an wie vielen Tagen Befragte in der vergangenen Woche mit einem Auto im jeweiligen Stadtgebiet gefahren sind („Gar nicht“, „An ein bis zwei Tagen“, „An drei bis vier Tagen“, „An fünf bis sechs Tagen“, „Täglich“), gemessen. Dabei wird jeweils der Mittelwert der Angabe (also etwa 1,5 Tage für „An ein bis zwei Tagen“) angesetzt.

bzw. die aufgrund einer Befreiung für Anwohner von der Zahlung von Mautgebühren ausgenommen wären, auf 0,01 € gesetzt. So können einerseits die Kosten einstellungskonformer Befürwortung auch weiterhin in logarithmierter Form in das Modell aufgenommen werden (siehe Gleichung (2.3)) und andererseits ist abgebildet, dass diesen Personen faktisch keine direkten Kosten in Form zu leistender Mautgebühren entstünden.

2.3.2. Analysemodell

Wie bereits erläutert, lässt sich anhand der logarithmierten Nachfragefunktion nach einstellungskonformem Handeln das funktionale Zusammenspiel der im Rahmen der LCH relevanten Variablen als Reihe linearer Beziehungen darstellen (siehe Gleichung (2.2)). Die postulierten Zusammenhänge sollen entsprechend anhand des linearen Regressionsmodells

$$\ln x_a = \text{const} + \gamma_\alpha \ln \alpha + \gamma_m \ln m + \gamma_p \ln p_a + \varepsilon \tag{2.3}$$

getestet werden. Ein strenger Test des Modells fordert, dass die Koeffizienten sich jeweils nicht signifikant von 1 (Umweltbewusstsein, Einkommen) oder -1 (Mautgebühren) unterscheiden (siehe Gleichung (2.2)). Wie auch Tutić et al. (2017) argumentieren, genügt in der Forschungspraxis ein moderaterer Test. So wird berücksichtigt, dass die Stärke der Einstellung entgegen der theoretischen Modellierung kein absolutes Skalenniveau aufweist. Zudem ist auf das Problem unbeobachteter Heterogenität hinzuweisen, das sich aus der der Nachfragefunktion nach einstellungskonformem Handeln inhärenten Annahme sonst gleicher Bedingungen ergibt. Während dem Problem unbeobachteter Heterogenität durch die randomisierte Zuweisung der Vignetten begegnet wird, kann insbesondere bei der vorliegenden Messung der Einstellung nicht von einem absoluten Skalenniveau ausgegangen werden. Es wird im Rahmen des moderaten Tests also lediglich geprüft, ob die Koeffizienten in die erwartete Richtung weisen, wobei $\gamma_\alpha > 0$, $\gamma_m > 0$ und $\gamma_p < 0$ jeweils statistisch signifikant sein muss.

2.4. Ergebnisse

Deskriptiv zeigt sich, dass der gesamte Wertebereich möglicher Mautbewertungen ausgeschöpft wird (siehe Abb. 2.3). Mit einem Mittelwert von 5,39 (Standardab-

weichung von 3,59) werden die Mautmodelle insgesamt eher kritisch bewertet. Das zeigt sich auch an dem vergleichsweise hohen Anteil stark ablehnender Mautbewertungen. So werden etwa 27% der vorgelegten Mautmodelle vollständig abgelehnt (Vignettenurteile = 1). Zugleich werden aber auch 41% der Mautmodelle eher oder stark befürwortet (Vignettenurteile > 6).

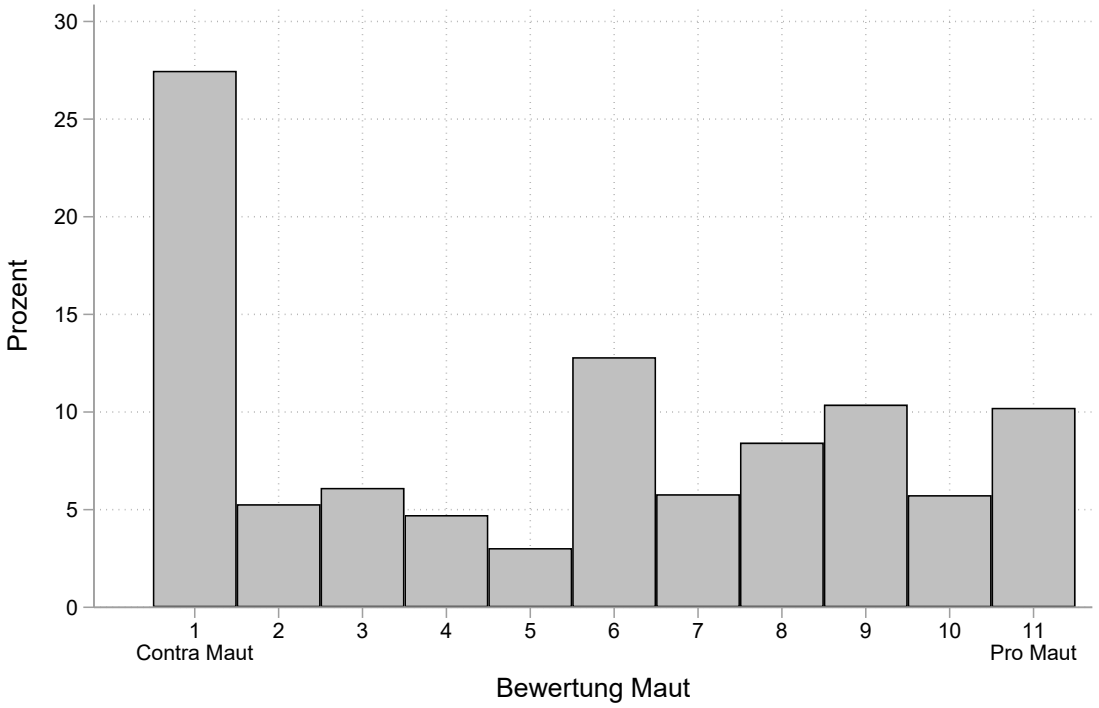

Abbildung 2.3.: Verteilung der Mautbewertungen (Darstellung basiert auf 4.317 Vignettenurteilen von 1.102 Personen)

2.4.1. Verbesserte Teststrategie zur Prüfung der LCH

Der Argumentation der LCH folgend sollte sich der negative Effekt der Höhe der Mautgebühren mit steigendem Umweltbewusstsein abschwächen. Um dies zu prüfen, wird das oben skizzierte Modell geschätzt (siehe Gleichung (2.3)), wobei die zentralen Variablen, wie ausgeführt, in logarithmierter Form eingehen. In Abb. 2.4 sind vorhergesagte Werte der logarithmierten Befürwortung einer Maut zunächst grafisch über das Umweltbewusstsein (links), die Höhe der Mautgebühren (mittig) und die Höhe des Haushaltsnettoäquivalenzeinkommens (rechts) dargestellt, wobei jeweils dieselben Referenzwerte wie in Abb. 2.1 eingesetzt sind.

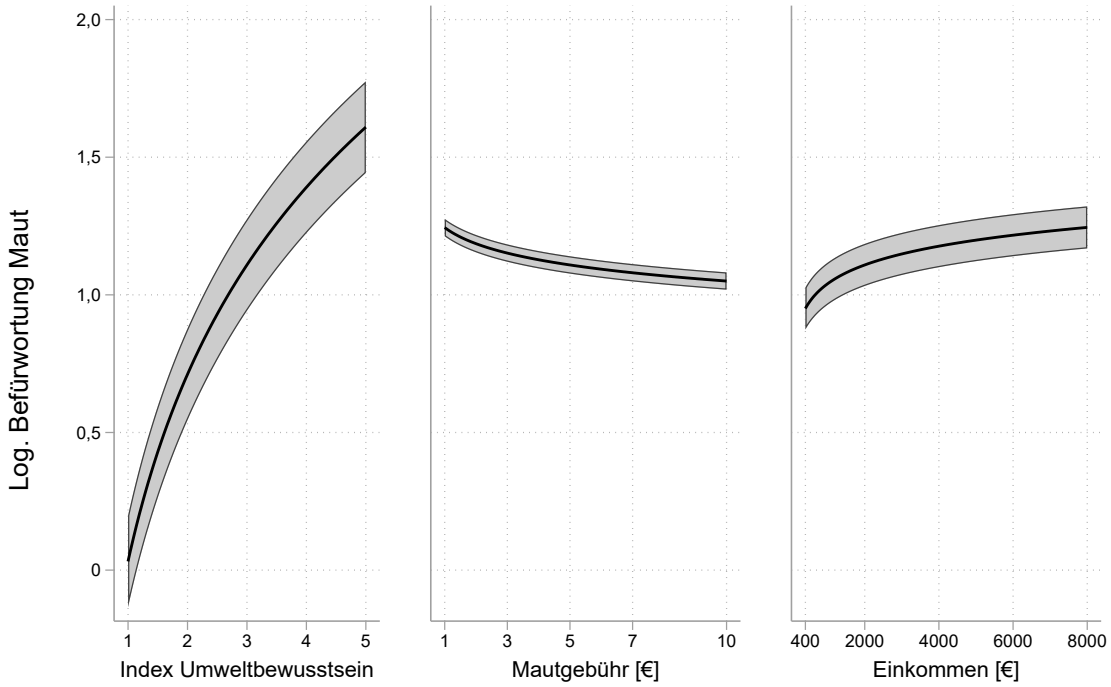

Abbildung 2.4.: Vorhergesagte Werte der logarithmierten Befürwortung einer Maut mit 95%-Konfidenzintervallen, dargestellt über das Umweltbewusstsein (links; wobei $m = 2000$, $p_a = 5$), die Mautgebühr (mittig; wobei $\alpha = 3$, $m = 2000$) und das Einkommen (rechts; wobei $\alpha = 3$, $p_a = 5$)

Es zeigt sich das erwartete Muster eines negativen Effekts der Höhe der Mautgebühr und positiver Effekte des Einkommens sowie des Umweltbewusstseins, die jeweils statistisch signifikant ($p < 0,05$) sind (siehe Tab. 2.1). Der Koeffizient des Umweltbewusstseins unterscheidet sich entsprechend der Forderungen eines strengen Tests zudem nicht signifikant vom Wert 1 (Wald-Test, $\mathrm{F}(1, 1101) = 0,05$, $p = 0,815$). Die übrigen Koeffizienten weisen zwar ebenfalls in die jeweils erwartete Richtung, unterscheiden sich jedoch von den anhand eines strengen Tests geforderten Werten von 1 für das Einkommen (Wald-Test, $\mathrm{F}(1, 1101) = 535,43$, $p < 0,001$) bzw. -1 für die Höhe der Mautgebühren (Wald-Test, $\mathrm{F}(1, 1101) = 3247,19$, $p < 0,001$).

2.4.2. Robustheitsanalysen

Um die Ergebnisse abzusichern, werden verschiedene Robustheitsanalysen durchgeführt. In einem ersten Schritt wird die Struktur der Daten (es liegen mehrere Vignet-

Tabelle 2.1.: Lineares Regressionsmodell zur Prüfung der Low-Cost-Hypothese

	Log. Befürwortung Maut
Log. Umweltbewusstsein	0,980***
	(0,085)
Log. Einkommen (in €)	0,098*
	(0,039)
Log. Mautgebühr (in €)	−0,084***
	(0,016)
Konstante	−0,580+
	(0,321)
Adj. R^2	0,066
N (Vignetten)	4317
N (Personen)	1102

Anmerkungen: Koeffizienten und cluster-robuste Standardfehler auf Befragtenebene in Klammern, um die Datenstruktur (mehrere Vignettenurteile pro Befragtem) zu berücksichtigen.
$^{+}p < 0{,}10$; $^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$

tenurteile pro Befragtem vor) den Empfehlungen zur Analyse faktorieller Surveys folgend durch die Verwendung eines Random Intercept (RI)-Modells explizit berücksichtigt (Auspurg und Hinz, 2015). Wie an den in Tab. 2.2 dargestellten Ergebnissen deutlich wird, weisen die Effekte erneut in die jeweils erwartete Richtung und sind statistisch signifikant ($p < 0,05$). Der Effekt des Umweltbewusstseins unterscheidet sich dabei zudem nicht signifikant von 1 (Wald-Test, $\chi^2(1) = 0,05$, $p = 0,831$), wohingegen sich die übrigen Koeffizienten von den im Rahmen eines strengen Tests geforderten Werten von 1 für das Einkommen (Wald-Test, $\chi^2(1) = 534,60$, $p < 0,001$) bzw. -1 für die Höhe der Mautgebühren (Wald-Test, $\chi^2(1) = 3976,66$, $p < 0,001$) unterscheiden. Insgesamt bleiben die berichteten Ergebnisse also auch unter Verwendung des RI-Modells robust. Darüber hinaus wird deutlich, dass mit rund 43% ein substanzieller Anteil der Varianz in den Vignettenurteilen auf Unterschiede zwischen den Befragten zurückführbar ist.

Vor diesem Hintergrund deutet die in Abb. 2.3 dargestellte Antwortverteilung darauf hin, dass es sich bei der Bewertung der Mautmodelle möglicherweise um einen zweiteiligen Entscheidungsprozess handelt, bei dem zunächst entschieden wird, ob eine Maut überhaupt in Erwägung gezogen wird und erst im zweiten Schritt (gegebenenfalls) eine Beurteilung des konkreten Modells vorgenommen wird. Sollte dies der Fall sein, könnten Schätzer einstufiger Regressionsverfahren verzerrt sein. Entsprechend wird hier zudem ein zweistufiges Vorgehen gewählt, das dem hohen Anteil

Tabelle 2.2.: Robustheitsanalysen zur Absicherung der Befunde zur Low-Cost-Hypothese anhand verschiedener Regressionsmodelle der logarithmierten Befürwortung einer City-Maut

	RI	Craggit[a]			OLS
		Stufe 1	Stufe 2	AMEs	
Log. Umweltbewusstsein	0,982***	1,334***	0,260***	0,960***	0,939***
	(0,085)	(0,140)	(0,060)	(0,059)	(0,084)
Log. Einkommen (in €)	0,098*	0,131*	0,030	0,098***	0,118**
	(0,039)	(0,062)	(0,022)	(0,022)	(0,039)
Log. Mautgebühr (in €)	−0,100***	−0,095***	−0,040***	−0,084***	–
	(0,014)	(0,025)	(0,010)	(0,014)	
Log. zu erwartende Kosten (in €)	–	–	–	–	−0,041***
					(0,007)
Konstante	−0,553+	−2,000***	1,320***	–	−0,934**
	(0,322)	(0,518)	(0,177)		(0,322)
σ[b]	–	0,491***	–	–	–
		(0,008)			
ρ[c]	0,426	–	–	–	–
Adj. R^2	0,066	–	–	–	0,074
N (Vignetten)	4317	4317		–	4317
N (Personen)	1102	1102		–	1102

Anmerkungen: Koeffizienten und cluster-robuste Standardfehler auf Befragtenebene in Klammern, um die Datenstruktur (mehrere Vignettenurteile pro Befragtem) zu berücksichtigen. Berichtet sind die Ergebnisse eines Random Intercept-Modells, eines Craggit-Modells (Cragg, 1971), sowie eines linearen Regressionsmodells mit einer alternativen Kostenmessung.
[a]Für Stufe 1 sind Koeffizienten eines Probitmodells, ob überhaupt ein Mautmodell in Erwägung gezogen wird ($y > 1$ vs. $y = 1$), dargestellt. Für Stufe 2 sind die Koeffizienten eines trunkierten linearen Regressionsmodells der Bewertung der Mautmodelle ($y > 1$) dargestellt. Die berichteten Average Marginal Effects (AMEs) fassen beide Stufen zusammen. Die Standardfehler der AMEs sind mithilfe des von Burke (2009) vorgeschlagenen Bootstrap-Verfahrens unter Verwendung des Stata ados craggit mit jeweils 100 Iterationen berechnet.
[b]σ gibt die geschätzte Fehlervarianz des Probitmodells an.
[c]ρ gibt die Intraclusterkorrelation des Random Intercept-Modells an.
$^{+}p < 0{,}10$; $^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$

stark negativer Bewertungen der Mautmodelle Rechnung trägt. Auf Stufe 1 wird dabei ein Probit-Modell geschätzt, wobei die abhängige Variable lediglich unterscheidet, ob eine Maut überhaupt in Erwägung gezogen wird (d.h. Abstufungen jenseits der Option „auf keinen Fall“ vorgenommen werden). Feingliedrige Abstufungen der Bewertung des jeweiligen Mautmodells gehen sodann auf Stufe 2 in ein trunkiertes lineares Regressionsmodell ein. Ergebnisse dieses sogenannten Craggit-

Modells (Cragg, 1971), das die Schätzungen beider Stufen kombiniert, sind ebenfalls in Tab. 2.2 dargestellt. Burke (2009) folgend, sind neben den getrennten Ergebnissen für beide Stufen auch durchschnittliche Marginaleffekte (AMEs) berechnet, die den Effekt der unabhängigen Variablen auf beiden Stufen zusammenfassen. Sie geben an, um wie viele Skalenpunkte ein Mautmodell besser bewertet wird, wenn sich die unabhängige Variable um eine Einheit erhöht. Entsprechende Standardfehler für die AMEs werden anhand des von Burke (2009) vorgeschlagenen Bootstrap-Verfahrens geschätzt.

Es zeigt sich, dass die Ergebnisse substanziell robust bleiben. Sowohl bei der Frage, ob eine Maut überhaupt in Erwägung gezogen wird (Stufe 1), als auch bei der Frage, in welchem Ausmaß ein solches Mautmodell befürwortet wird (Stufe 2), weisen die Effekte in die erwartete Richtung. Selbiges gilt für die durchschnittlichen Marginaleffekte, die beide Stufen zusammenfassen. Die AMEs sind jeweils statistisch signifikant ($p < 0,001$), unterscheiden sich jedoch auch signifikant von den im Rahmen eines strengen Tests geforderten Werten von -1 für die Höhe der Mautgebühren (Wald-Test, $\chi^2(1) = 3332,76$, $p < 0,001$) bzw. 1 für die Höhe des Einkommens (Wald-Test, $\chi^2(1) = 1284,65$, $p < 0,001$). Lediglich der durchschnittliche Marginaleffekt der Stärke des Umweltbewusstseins unterscheidet sich nicht signifikant von 1 (Wald-Test, $\chi^2(1) = 0,46$, $p = 0,497$), wie es ein strenger Test erfordert. Dennoch erfüllen die Ergebnisse beider Modellierungen die Anforderungen eines moderaten Tests, wie er angesichts der bereits ausgeführten Beschränkungen in der Forschungspraxis angemessen scheint.

In einem dritten Schritt soll geprüft werden, inwiefern die vermuteten Zusammenhänge im Speziellen für Personen bestehen, die in ihrem jeweiligen Mobilitätsverhalten direkt betroffen wären. Hierzu wird eine alternative Operationalisierung der individuellen Kosten einer City-Maut verwendet, die auf die tatsächlich zu erwartende Höhe anfallender Mautgebühren bei gegebenem Mobilitätsverhalten abzielt (siehe Abschnitt 2.3.1). Wie aus Tab. 2.2 ersichtlich ist, bleibt der Kernbefund auch unter Verwendung dieser alternativen Operationalisierung der Kosten einstellungskonformer Befürwortung bestehen.[10]

10 Dieser Befund bleibt (für beide Kostenmessungen) auch dann bestehen, wenn im Rahmen weiterer Robustheitsanalysen sehr hohe Einkommen (über 8.000€ Haushaltsnettoäquivalenzeinkommen) ausgeschlossen werden.

2.4.3. Prüfung der LCH anhand der vormals üblichen Teststrategie

Schließlich wird untersucht, inwiefern die vormals gängige Teststrategie, die LCH mithilfe eines Interaktionsterms aus Einstellung und Kosten zu prüfen, zu substanziell vergleichbaren Ergebnissen führt. Hierzu wird also neben den Haupteffekten des Umweltbewusstseins und der Höhe der Mautgebühren ein multiplikativer Term in das Modell aufgenommen (siehe Tab. 2.3). Dabei wird neben der verbreiteten OLS-Modellierung auch die bereits angesprochene RI-Modellierung verwendet, um die genestete Struktur der Daten explizit zu berücksichtigen. Zur besseren Vergleichbarkeit ist wie auch in den vorigen Abschnitten die abhängige Variable der Befürwortung einer Maut jeweils logarithmiert, wobei die dargestellten Ergebnisse auch unter Verwendung der nicht-logarithmierten Form robust bleiben. Erwartungsgemäß zeigt sich in allen vier Modellen ein positiver Effekt des Umweltbewusstseins ($p < 0,05$) und ein negativer Effekt der Höhe der Mautgebühren ($p < 0,05$). Der im Rahmen der LCH zentrale Interaktionseffekt aus Einstellung und den Kosten einstellungskonformen Handelns kann jedoch nicht beobachtet werden. Zwar ist der Interaktionseffekt aus Umweltbewusstsein und Höhe der Mautgebühr im RI-Modell statistisch signifikant ($p < 0,05$), weist jedoch entgegen der theoretischen Erwartung kein negatives Vorzeichen auf.

Während die LCH anhand der vormals gängigen Teststrategie verworfen würde, liefert die verbesserte Teststrategie durchweg Unterstützung für die LCH. Die verbesserte Teststrategie scheint, zumindest basierend auf den vorliegenden Ergebnissen, eine (minimal) höhere Erklärungskraft aufzuweisen und zudem robuster gegenüber einzelnen Modellierungsentscheidungen (wie der konkreten Operationalisierung der Kosten einstellungskonformen Verhaltens) zu sein.

2.5. Zusammenfassung

Die vorliegende Arbeit zielte auf einen Test der Implikationen, die sich aus der von Tutić et al. (2017) vorgeschlagenen neuerlichen Herleitung der LCH aus einem mikroökonomischen Modell ergeben, anhand der Befürwortung einer City-Maut in München. Dabei sollte in zweierlei Hinsicht an die umfassende Literatur um die LCH angeknüpft werden. Erstens sollte die LCH im engeren Sinn eines Preiseffekts, wie sie sich aus der verbesserten Teststrategie ergibt, einem ersten empirischen Test un-

Tabelle 2.3.: Lineare Regressionsmodelle der logarithmierten Befürwortung einer City-Maut zur Prüfung der Low-Cost-Hypothese anhand der vormals üblichen Teststrategie

	OLS		RI	
	Ia	IIa	Ib	IIb
Umweltbewusstsein	0,233***	0,267***	0,203***	0,270***
	(0,042)	(0,029)	(0,039)	(0,029)
Mautgebühr (in €)	−0,058*	–	−0,084***	–
	(0,023)		(0,019)	
Umweltbewusstsein × Mautgebühr	0,009	–	0,015**	–
	(0,006)		(0,005)	
Zu erwartende Kosten (in €)	–	−0,015**	–	−0,013*
		(0,006)		(0,005)
Umweltbewusstsein × zu erwartende Kosten	–	0,002	–	0,001
		(0,002)		(0,001)
Konstante	0,581***	0,364**	0,725***	0,366**
	(0,165)	(0,114)	(0,152)	(0,112)
ρ[a]	–	–	0,433	0,433
Adj. R^2	0,058	0,062	0,058	0,061
N (Vignetten)	4317	4317	4317	4317
N (Personen)	1102	1102	1102	1102

Anmerkungen: Koeffizienten und cluster-robuste Standardfehler auf Befragtenebene in Klammern, um die Datenstruktur (mehrere Vignettenurteile pro Befragtem) zu berücksichtigen. Berichtet sind Ergebnisse für die Operationalisierung der Kosten anhand der Mautgebühr (Ia und Ib) bzw. der alternativen Messung anhand der zu erwartenden Kosten (IIa und IIb) jeweils unter Verwendung eines linearen Regressionsmodells bzw. eines Random-Intercept-Modells. Die abhängige Variable ist zur besseren Vergleichbarkeit mit den in Abschnitt 2.4.1 und 2.4.2 berichteten Modellen jeweils logarithmiert. Die dargestellten Effekte bleiben substanziell aber auch unter Verwendung der nichtlogarithmierten Form und unter Ausschluss sehr hoher Einkommen (über 8000 € Haushaltsnettoäquivalenzeinkommen) erhalten.
[a] ρ gibt die Intraclusterkorrelation des Random-Intercept-Modells an.
$^{+}p < 0{,}10$; $^{*}p < 0{,}05$; $^{**}p < 0{,}01$; $^{***}p < 0{,}001$

terzogen und so die Ausführungen von Tutić et al. (2017) um einen zentralen Aspekt ergänzt werden. Zweitens sollte anhand des gewählten Anwendungsfalls untersucht werden, inwiefern bereits bei der vorgelagerten Entscheidung über die Gestaltung gesetzlicher Rahmenbedingungen trotz individuell anfallender Kosten Bereitschaft zu einstellungskonformer Befürwortung einer City-Maut besteht.

Den vorliegenden Ansatz zeichnet dabei aus, dass das experimentelle Design eines faktoriellen Surveys im Rahmen einer Bevölkerungsbefragung für den empirischen Test der LCH herangezogen wurde. Über 1.300 Befragte aus München und eini-

ger Umlandgemeinden mit hoher Pendelverflechtung zu München bewerteten insgesamt rund 5.000 fiktive Mautmodelle. Aufgrund der randomisierten Zuweisung der verschiedenen Mautmodelle (und damit der experimentellen Stimuli) konnte ausgeschlossen werden, dass Charakteristika der vorgelegten Mautmodelle mit soziodemographischen Merkmalen der Befragten korrelieren. So war es einerseits möglich, Bevölkerungsgruppen einzubeziehen, die in unterschiedlichem Maße von (den Kosten) einer möglichen City-Maut betroffen wären und so die sich aus der LCH ergebenden Implikationen anhand einer umfassenden Datengrundlage zu testen. Andererseits konnte die Höhe der Mautgebühren und damit die Kostspieligkeit einer möglichen City-Maut tatsächlich exogen variiert werden.

Empirisch zeigte sich eine breite Streuung der Bewertungen über alle vorgelegten Mautmodelle hinweg. Bezogen auf die LCH fanden sich durchaus Hinweise, die den erwarteten Zusammenhängen entsprechen. Während ein stärker ausgeprägtes Umweltbewusstsein sowie ein höheres Einkommen zu positiveren Bewertungen der Mautmodelle führten, sank die Zustimmung mit steigenden Mautgebühren. Die Effekte wiesen also in die erwartete Richtung und waren zudem statistisch signifikant. Im Rahmen weitergehender Analysen (Modellierung mithilfe eines Random Intercept-Modells sowie eines Craggit-Modells, Ausschluss von Befragten mit sehr hohem Einkommen, Verwendung einer alternativen Operationalisierung der Kosten) erwies sich dieser Befund als sehr robust.

Abschließend wurden diese Ergebnisse analog zum Vorgehen von Tutić et al. (2017) mit auf der vormals gängigen Teststrategie basierenden Berechnungen verglichen. Letztere lieferten insgesamt keine Unterstützung für die LCH. Der im Fokus stehende Interaktionseffekt aus Einstellungs- und Kostenvariable wies entgegen der theoretischen Erwartung kein negatives Vorzeichen auf und reagierte in Stärke und Signifikanz auf einzelne Modellierungsentscheidungen. Neben der umfassenderen theoretischen Fundierung scheint die verbesserte Teststrategie also auch zu empirisch zuverlässigeren Resultaten zu führen, die zudem eine (wenngleich nur minimal) höhere Erklärungskraft aufwiesen.

2.6. Diskussion

Die vorliegende Analyse liefert einen wichtigen Beitrag zur Literatur um die LCH, indem die von Tutić et al. (2017) vorgeschlagene neuerliche Herleitung der LCH (im engeren Sinn eines Preiseffekts) einem ersten empirischen Test unterzogen wurde.

Dazu wurden die Vorteile eines faktoriellen Surveys im Rahmen einer Bevölkerungsbefragung genutzt. Dieses experimentelle Design erlaubt die Prüfung der theoretisch vermuteten kausalen Zusammenhänge, die sich aus der LCH ergeben. Durch den breiten Einbezug verschiedener Aspekte einer möglichen City-Maut wurden den Befragten im Vergleich zu einfachen Itemabfragen komplexere, zugleich aber auch anschaulichere Stimuli vorgelegt (vgl. Auspurg und Liebe, 2011).

Dabei wurde eine für faktorielle Surveys typische elfstufige Antwortskala verwendet. Dagegen kann eingewandt werden, dass dieses Vorgehen im Vergleich zu einem Choice-Experiment weiter von einer realitätsnahen Entscheidungssituation entfernt ist, wie sie etwa in einer möglichen Abstimmung über ein oder zwei konkrete Mautmodelle vorkommt (vgl. etwa Auspurg und Liebe, 2011, S. 304). Allerdings mag die Möglichkeit feingliedriger Abstufungen des Vignettenurteils gerade in Situationen mit einer Vielzahl kritischer Bewertungen sinnvoll sein. Immerhin sind vorgenommene Abwägungen so auch dann beobachtbar, wenn Befragte dem vorgelegten Mautmodell insgesamt eher kritisch gegenüberstehen.

Die exogene Variation u.a. der Höhe der Mautgebühren stellt einen wesentlichen Vorteil des gewählten experimentellen Designs dar. Dennoch muss darauf hingewiesen werden, dass aus den exogen variierten Mautgebühren nur eingeschränkt auf die tatsächlich mit einer möglichen City-Maut einhergehenden Kosten geschlossen werden kann. Immerhin könnte durch den Umstieg auf andere Verkehrsmittel oder eine grundsätzliche Anpassung des eigenen Mobilitätsverhaltens (Vermeidung unnötiger Wegstrecken) die individuelle Betroffenheit von Mautgebühren vermieden werden (darauf deuten etwa auch Befunde zum Verkehrsaufkommen nach Einführung einer City-Maut in London, siehe etwa Ellison et al., 2013). Fasst man die daraus resultierenden Komfortverluste jedoch als nicht-monetäre Kostenkomponente auf (vgl. bspw. Brüderl und Preisendörfer, 1995), beschneidet dies die Aussagekraft der dargelegten Befunde nicht.

Die zentrale Einschränkung bezieht sich vielmehr darauf, dass es sich bei der Bewertung der vorgelegten Mautmodelle um hypothetische und keine realen Entscheidungen handelte. Die untersuchte abhängige Variable beschreibt die Verhaltensabsicht, eine City-Maut befürworten zu wollen, nicht aber die tatsächliche Befürwortung. Mithin ist fraglich, inwiefern sich eine solche Absicht auch auf tatsächliches Entscheidungsverhalten übertragen lässt. Empirische Untersuchungen zur Validität der in Surveyexperimenten gemessenen Verhaltensabsichten in Bezug auf tatsächliches Verhalten haben teils widersprüchliche Befunde hervorgebracht. Für die

vorliegende Untersuchung zentral scheint die sich abzeichnende Erkenntnis, dass in Situationen, in denen kaum normative Erwartungen eines bestimmten Verhaltens bestehen, von einer hohen Übereinstimmung zwischen Verhaltensabsicht und tatsächlichem Verhalten ausgegangen werden kann – zumindest hinsichtlich der Richtung und relativen Stärke von Effekten (vgl. Petzold und Wolbring, 2019, S. 8). Der möglicherweise dennoch mangelnden externen Validität des Ansatzes sollte durch eine realitätsnahe Ausgestaltung der vorgelegten Beschreibungen möglicher Mautmodelle begegnet werden. Auch die Bewertung verschiedener Mautmodelle, bevor durch die Einführung einer City-Maut tatsächlich Kosten in Form zu leistender Mautgebühren anfallen, korrespondiert mit realen Prozessen politischer Partizipation. Nichtsdestotrotz wären die hier anhand von Verhaltensabsichten untersuchten Zusammenhänge in künftigen empirischen Arbeiten auch anhand echter Verhaltensdaten zu prüfen, um die externe Validität der Befunde abzusichern. Zu denken sei dabei beispielsweise an die Untersuchung einer politisch verbindlichen Bürgerbefragung. So könnte anstatt der geäußerten Bereitschaft, das jeweilige Mautmodell zu befürworten, die tatsächliche Befürwortung einer City-Maut in einer Entscheidungssituation untersucht werden, aus der reale Kosten erwachsen (können).

Solange die für reale (Verhaltens-)Entscheidungen relevanten Faktoren nicht mit den hier untersuchten interagieren, sollte aber zumindest die hohe interne Validität des experimentellen Designs gewährleistet bleiben (vgl. Auspurg et al., 2014, S. 44). Zumal es keine Hinweise darauf gibt, dass die in realen Entscheidungen relevanten Faktoren sich derart von den hier untersuchten unterscheiden, bleibt der Kernbefund der vorliegenden Analyse bestehen: Anhand der mikroökonomisch fundierten Teststrategie, wie sie Tutić et al. (2017) vorschlagen, wird die LCH am Beispiel der Befürwortung einer City-Maut in München gestützt.

Während sich die auf der verbesserten Teststrategie basierenden Ergebnisse (trotz nur geringfügig höherer Erklärungskraft der Modelle) als sehr robust erweisen, scheinen die auf der vormals gängigen Teststrategie beruhenden Ergebnisse deutlich auf einzelne Modellierungsentscheidungen zu reagieren. Der viel gravierendere Unterschied zwischen den beiden Teststrategien besteht jedoch darin, dass sie zu inhaltlich unterschiedlichen Schlussfolgerungen führen. Auf Basis der herkömmlichen Teststrategie würde man die LCH verwerfen, obwohl sie sich anhand der verbesserten Teststrategie bewährt.

Insofern stehen die hier vorgelegten empirischen Befunde auch nicht im Widerspruch zu den Ausführungen von Tutić et al. (2017), sondern ergänzen einerseits

den noch ausstehenden Test der LCH im engeren Sinn eines Preiseffekts einstellungskonformen Handelns und deuten andererseits einmal mehr darauf hin, dass die theoretisch fundierte Teststrategie dem vormals gängigen Vorgehen auch empirisch vorzuziehen scheint. Es ist jedoch darauf hinzuweisen, dass es sich hierbei um keinen systematischen Vergleich der beiden Teststrategien hinsichtlich ihres generellen empirischen Abschneidens handelt. Eine solche Gegenüberstellung wäre im Rahmen künftiger Forschung zu leisten.

Darüber hinaus scheinen die vorgelegten Befunde auch für die umweltpolitische Debatte um die Konzeption wirksamer Maßnahmen, etwa zur Verringerung der (straßenverkehrsbedingten) Luftschadstoffbelastung in Ballungsräumen, relevant. Neben der breiten Streuung in der Bewertung der hypothetischen Mautmodelle zeigt sich auch das im Rahmen der LCH postulierte Zusammenspiel von Einstellung und Kosten. Aus praktischer Perspektive verweist dies einerseits auf die Bedeutung der langfristigen Entwicklung des Umweltbewusstseins (bspw. Franzen und Vogl, 2013) und rückt andererseits die Kostenkomponente als möglichen Ansatzpunkt in den Fokus. Aufschlussreich ist dabei, dass sich die vermuteten Zusammenhänge bereits bei der vorgelagerten Bereitschaft zur einstellungskonformen Veränderung (gesetzlicher) Rahmenbedingungen zeigen – noch bevor also tatsächliche Kosten erwachsen. Das mag für die Konzeption (umwelt-)politischer Maßnahmenbündel bedeutend sein, die sowohl öffentlichen Zuspruch erfahren als auch steuerungspolitische Wirkung entfalten (bspw. Wicki et al., 2019). Dabei ist im Einzelfall auch das Angebot möglicher Handlungsalternativen, inwiefern etwa die (wahrgenommene) Möglichkeit zum Umstieg auf andere Verkehrsmittel besteht, zu berücksichtigen.

Danksagung

Wertvolle Anmerkungen zu einer früheren Version des Manuskripts sind Katrin Auspurg, Christian Ganser und Werner Fröhlich sowie zwei anonymen Gutachterinnen oder Gutachtern und der Herausgeberin und den Herausgebern der KZfSS zu verdanken. Ulf Liebe sei für die freundliche Bereitstellung der Analysefiles zu Tutić et al. (2017) gedankt.

Data note

Die Studie wurde zusammen mit Katrin Auspurg und Sabine Düval durchgeführt. Für Unterstützung bei der Umsetzung sei zudem den Kolleginnen und Kollegen Christiane Bozoyan, Werner Fröhlich, Christian Ganser, Bettina Pettinger und Michael Rochlitz sowie allen beteiligten studentischen Hilfskräften gedankt. Den Teilnehmerinnen und Teilnehmern der Methoden-Kurse im Wintersemester 2017/18 sei für ihre Mitarbeit an der Konzeption des Fragebogens gedankt. Die Erhebung fand zwischen Mitte Mai und Anfang Juli 2018 statt, wobei einzelne Fragebögen noch bis Ende Juli nacherfasst wurden. Die Studie wurde aus Institutsmitteln finanziert. Analysefiles sind auf Anfrage erhältlich.

Literaturverzeichnis

Apte, Joshua S., Michael Brauer, Aaron J. Cohen, Majid Ezzati, und C. Arden Pope. 2018. „Ambient $PM_{2.5}$ Reduces Global and Regional Life Expectancy." *Environmental Science & Technology Letters* 5:546–551.

Auspurg, Katrin, Corinna Frodermann, und Thomas Hinz. 2014. „Berufliche Umzugsentscheidungen in Partnerschaften. Eine experimentelle Prüfung von Verhandlungstheorie, Frame-Selektion und Low-Cost-These." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 66:21–50.

Auspurg, Katrin und Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, California: SAGE Publications.

Auspurg, Katrin und Ulf Liebe. 2011. „Choice-Experimente und die Messung von Handlungsentscheidungen in der Soziologie." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 63:301–314.

Best, Henning. 2008. „Die Umstellung auf ökologische Landwirtschaft. Empirische Analysen zur Low-Cost-Hypothese des Umweltverhaltens." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 60:314–338.

Best, Henning. 2009a. „Kommt erst das Fressen und dann die Moral? Eine feldexperimentelle Überprüfung der Low-Cost-Hypothese und des Modells der Frame-Selektion." *Zeitschrift für Soziologie* 38:131–151.

Best, Henning. 2009b. „Structural and Ideological Determinants of Household Waste Recycling: Results from an Empirical Study in Cologne, Germany." *Nature and Culture* 4:167–190.

Best, Henning und Thorsten Kneip. 2011. „The Impact of Attitudes and Behavioral Costs on Environmental Behavior: A Natural Experiment on Household Waste Recycling." *Social Science Research* 40:917–930.

Best, Henning und Clemens Kroneberg. 2012. „Die Low-Cost-Hypothese. Theoretische Grundlagen und empirische Implikationen.“ *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 64:535–561.

Braun, Norman und Axel Franzen. 1995. „Rationalität und Umweltverhalten.“ *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 47:231–248.

Braun, Norman und Thomas Gautschi. 2011. *Rational-Choice-Theorie.* Weinheim: Juventa.

Brüderl, Josef und Peter Preisendörfer. 1995. „Der Weg zum Arbeitsplatz: Eine empirische Untersuchung zur Verkehrsmittelwahl“, S. 69–88. In: Diekmann, Andreas und Axel Franzen (Hrsg.): *Kooperatives Umwelthandeln: Modelle, Erfahrungen, Massnahmen.* Chur/Zürich: Verlag Rüegger AG.

Burke, William J. 2009. „Fitting and Interpreting Cragg's Tobit Alternative using Stata.“ *The Stata Journal* 9:584–592.

Cragg, John G. 1971. „Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods.“ *Econometrica* 39:829–844.

Derksen, Linda und John Gartrell. 1993. „The Social Context of Recycling.“ *American Sociological Review* 58:434–442.

Diekmann, Andreas und Peter Preisendörfer. 1998. „Umweltbewußtsein und Umweltverhalten in Low- und High-Cost-Situationen. Eine empirische Überprüfung der Low-Cost-Hypothese.“ *Zeitschrift für Soziologie* 27:438–453.

Diekmann, Andreas und Peter Preisendörfer. 2003. „Green and Greenback. The behavioral effects of environmental attitudes in low-cost and high-cost situations.“ *Rationality and Society* 15:441–472.

Diekmann, Jochen. 1998. „Umwelt, Ökonomik und empirische Sozialforschung. Bemerkungen zum interdisziplinären Diskurs.“, S. 187–198. In: Schupp, Jürgen und Gert Wagner (Hrsg.): *Umwelt und empirische Sozial- und Wirtschaftsforschung.* Berlin: Duncker und Humblot.

Ellison, Richard B., Stephen P. Greaves, und David A. Hensher. 2013. „Five Years of London's Low Emission Zone: Effects on Vehicle Fleet Composition and Air Quality.“ *Transport Research Part D* 23:25–33.

Findley, Michael G., Brock Laney, Daniel L. Nielson, und J. C. Sharman. 2017. „External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation." *The Journal of Politics* 79:856–872.

Franzen, Axel und Dominikus Vogl. 2013. „Two decades of measuring environmental attitudes: A comparative analysis of 33 countries." *Global Environmental Change* 23:1001–1008.

Hainmueller, Jens, Dominik Hangartner, und Teppei Yamamoto. 2015. „Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112:2395.

Keuschnigg, Marc und Fabian Kratz. 2018. „Thou Shalt Recycle: How Social Norms of Environmental Protection Narrow the Scope of the Low-Cost Hypothesis." *Environment and Behavior* 50:1059–1091.

Lelieveld, Jos, Klaus Klingmüller, Andrea Pozzer, Ulrich Pöschl, Mohammed Fnais, Andreas Daiber, und Thomas Münzel. 2019. „Cardiovascular Disease Burden from Ambient Air Pollution in Europe Reassessed Using Novel Hazard Ratio Functions." *European Heart Journal* 40:1590–1596.

Liebe, Ulf. 2007. *Zahlungsbereitschaft für kollektive Umweltgüter. Soziologische und ökonomische Analysen.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Liebe, Ulf und Benedikt Jahnke. 2017. „Giving More to Humans than to Animals in Need? A Behavioral Measure of Animal-Human Continuity in Large-scale Surveys." *Journal of International Society for Anthrozoology* 30:249–262.

Mas-Collel, Andreu, Michael D. Winston, und Jerry R. Green. 1995. *Microeconomic theory*. New York: Oxford University Press.

Mayerl, Jochen. 2010. „Die Low-Cost-Hypothese ist nicht genug. Eine Empirische Überprüfung von Varianten des Modells der Frame-Selektion zur besseren Vorhersage der Einflussstärke von Einstellungen auf Verhalten." *Zeitschrift für Soziologie* 39:38–59.

Neumann, Robert und Guido Mehlkop. 2018. „Umweltentscheidungen als Wechselspiel von Einstellungen, Handlungskosten und situativer Rahmung – ein empirischer Theorienvergleich mit Daten des GESIS Panels." *Zeitschrift für Soziologie* 47:101–118.

Petzold, Knut und Tobias Wolbring. 2019. „What Can We Learn from Factorial Surveys About Human Behavior? A Validation Study Comparing Field and Survey Experiments on Discrimination." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 15:19–30.

Pope, C. Arden und Douglas W. Dockery. 2006. „Health Effects of Fine Particulate Air Pollution: Lines that Connect." *Journal of Air & Waste Management Association* 56:709–742.

Preisendörfer, Peter. 2000. „Strukturell-situationale Gegebenheiten als Bestimmungsfaktoren der Verkehrsmittelwahl." *Soziale Welt* 51:487–501.

Rauhut, Heiko und Ivar Krumpal. 2008. „Die Durchsetzung sozialer Normen in Low-Cost und High-Cost Situationen." *Zeitschrift für Soziologie* 37:380–402.

Schultz, P. Wesley und Stuart Oskamp. 1996. „Effort as a Moderator of the Attitude-Behavior Relationship: General Environmental Concern and Recycling." *Social Psychology Quarterly* 59:375–383.

Tutić, Andreas, Thomas Voss, und Ulf Liebe. 2017. „Low-Cost-Hypothese und Rationalität. Eine neue theoretische Herleitung und einige Implikationen." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69:651–672.

Umweltbundesamt. 2017. *Daten zur Umwelt 2017: Indikatorenbericht*. Dessau-Roßlau: Umweltbundesamt.

Varian, Hal R. 1992. *Microeconomic analysis*. New York: Norton.

Wicki, Michael, Lukas Fesenfeld, und Thomas Bernauer. 2019. „In search of politically feasible policy-packages for sustainable passenger transport: Insights from choice experiments in China, Germany, and the USA." *Environmental Research Letters* 14:084048.

2.A. Anhang

Tabelle 2.A1.: Übersicht der im Rahmen der Low-Cost-Hypothese vermuteten Zusammenhänge anhand partieller Ableitungen der Nachfragefunktion (vgl. Tutić et al., 2017, S. 657)

Vermuteter Zusammenhang	Partielle Ableitung
Positiver Effekt der Einstellung	$\frac{\partial x_a^*(\alpha,m,p_a)}{\partial \alpha} = \frac{m}{p_a} > 0$
Negativer Effekt der Kosten einstellungskonformen Handelns	$\frac{\partial x_a^*(\alpha,m,p_a)}{\partial p_a} = -\alpha\frac{m}{p_a^2} < 0$
Kostenabhängige Stärke des Einstellungseffekts	$\frac{\partial^2 x_a^*(\alpha,m,p_a)}{\partial p_a \partial \alpha} = \frac{\partial \frac{m}{p_a}}{\partial p_a} = -\frac{m}{p_a^2} < 0$

Tabelle 2.A2.: Dimensionen und Levels der fiktiven Mautmodelle

	Dimension	Level	Σ
1	Verbesserung der Luftqualität im Stadtgebiet München[a]	• Um 1% • [Um 5%] • Um 10% • [Um 15%] • Um 20%	5
2	Orte mit besonders starker Luftverbesserung	• Gesamte Mautzone • Übermäßig belastete Kreuzungen • Keine (gleichmäßige Verbesserung im Stadtgebiet)	3
3	Verkehrsaufkommen im Stadtgebiet	• Unverändert • Etwas geringer • Deutlich geringer	3
4	Geltungsbereich der Maut	• Innerhalb des Altstadtrings (mit Altstadtring) • Innerhalb des Mittleren Rings (ohne Ring) • Innerhalb des Mittleren Rings (mit Ring) • Gesamtes Stadtgebiet (ohne Autobahn)	4
5	Von Maut betroffene Fahrzeuge	• Alle Fahrzeuge • Moderat und sehr schadstoffreiche Fahrzeuge (alle ohne „blaue Plakette“) • Sehr schadstoffreiche Fahrzeuge (alle ohne „grüne Plakette“)	3
6	Mautgebühren pro Tag[a]	• 1 € • [3 €] • 5 € • [7 €] • 10 €	5
7	Verwendung der Mautgebühren	• Förderung des öffentlichen Nahverkehrs • Schaffung von Grünflächen im Stadtgebiet • Baumaßnahmen zur Stauvermeidung	3
8	Mautbefreiung für Anwohner	• Ja • Nein	2
Σ	Vignettenuniversum ($5 \times 3 \times 3 \times 4 \times 3 \times 5 \times 3 \times 2$)		16.200

Anmerkungen: [a]Grad der Verbesserung der Luftqualität im Stadtgebiet sowie die Höhe der Maut wurde jeweils bei der Hälfte der Befragten zwischen 3 und bei der anderen Hälfte zwischen 5 möglichen Ausprägungen variiert, um mögliche nichtlineare Effekte besser schätzen zu können (betreffende Ausprägungen in eckigen Klammern).

3. Support for city road tolls: a question of self-interest?

Dieses Kapitel ist erschienen als:

Thiel, Fabian. 2021. „Support for city road tolls: a question of self-interest?“ In: Franzen, Axel und Sebastian Mader (Hrsg.): *Research Handbook on Environmental Sociology*. Cheltenham, UK: Edward Elgar Publishing. DOI: https://doi.org/10.4337/9781800370456.00026

Zusammenfassung

Ziel der vorliegenden Studie ist zu einem klareren Verständnis der Faktoren, die Einstellungen gegenüber City-Mautgebühren beeinflussen, beizutragen. Zieht sich die berichtete insgesamt geringe Befürwortung durch die gesamte Bevölkerung? Die präsentierten Analysen basieren auf den Daten einer Bevölkerungsbefragung in München und vier Umlandgemeinden. Mehr als 1.300 Befragte bewerteten im Frühsommer 2018 über 5.300 fiktive Mautmodelle. Bestehende Ansätze werden dabei durch die einzigartige Kombination aus faktoriellem Surveyexperiment mit Informationen zur persönlichen Fahrzeugnutzung sowie zum Wohnumfeld der Befragten erweitert. So lassen sich durch die Berücksichtigung spezifischer Subgruppen, die wohl in unterschiedlichem Maße von der Einführung einer City-Maut betroffen wären, zugrunde liegende Mechanismen entflechten. Die Ergebnisse zeigen, dass 1) die Befürwortung insgesamt gering ist, jedoch zwischen den Subgruppen variiert, dass 2) die institutionelle Ausgestaltung zwar insgesamt wichtig erscheint, allerdings 3) relevante Faktoren sich ebenfalls zwischen Subgruppen unterscheiden. Neben allgemeinen politischen Überzeugungen erweist sich insbesondere individuelles Eigeninteresse als wesentlich bei der Entstehung von Einstellungen gegenüber einer City-Maut.

Abstract

This study aims to more clearly understand which factors shape attitudes toward city road tolls. Does the reported overall low support permeate the entire population? Presented analyses are based on population survey data from Munich and four surrounding municipalities. More than 1,300 respondents rated more than 5,300 fictious toll schemes in the early summer of 2018. A unique combination of a factorial survey experiment, information on personal vehicle usage, and neighborhood context information extends existing approaches. By distinguishing specific subgroups possibly affected differently by the introduction of a city road toll, underlying mechanisms can be disentangled. Results show that 1) overall support is low, but differs between subgroups, that 2) institutional design is important to citizens, but that 3) relevant factors also differ between subgroups. Aside from general beliefs concerning policy consequences, individual self-interest proves particularly important in the formation of attitudes toward city road tolls.

3.1. Urban road pricing

Traffic jams and noise as well as environmental consequences (e.g., polluted air) are characteristic of urban mobility in many places. Inefficient traffic flow not only causes time loss for commuters (e.g., Santos and Shaffer, 2004), but high traffic volume also contributes to noise and air pollution, which in turn affect public health in the area (e.g., Lee, 2018; Pope and Dockery, 2006). In some cities, existing measures to reduce the negative impacts of road traffic (e.g., low emission zones) seem to mitigate the issues to some extent (Ellison et al., 2013; Fensterer et al., 2014; Holman et al., 2015; Qadir et al., 2013). Their effectiveness in reducing concentrations of air pollution is, however, low (Morfeld et al., 2014b,a, 2015). Moreover, given the persistent traffic obstacles and exceedance of air pollutant limit values in cities with existing low emission zones, it would appear these measures are not sufficient.

Despite some technical achievements, such as the development of more fuel-efficient vehicles that emit fewer air pollutants as well as the pronounced environmental concern observed over recent decades (Franzen and Vogl, 2013), the problems associated with road traffic, especially in large cities, remain unresolved. The dilemma structure of transport mode choice can help to explain this stagnation. Santos and Shaffer (2004) describe this as the modern version of Hardin's 1968 "tragedy of the commons": While each person weighs the individual (marginal) costs against the benefits of various possible means of transportation when faced with a decision, they typically do not consider the external costs that arise from each of the options. This "free-rider" phenomenon results in the over-usage of collective goods such as a city's road network, which results in the associated negative consequences (i.e., congestion, noise, and air pollution).

A widely-discussed institutional solution to this dilemma would be the introduction of a congestion tax. By internalizing the externally generated costs of mobility behavior, such a tax would alter the incentive structure for individual commuters and thus counteract the over-usage of the public good. Optimal pricing levels for the internalization of externalities could be achieved by using marginal cost pricing, as proposed by Pigou (1929) and subsequently taken up by, for example, Walters (1961), who applied the concept on pricing the costs of highway congestion. Essentially, the idea is to charge a fee or tax on each unit of road usage that is exactly equal to the externally generated extra cost of this additional unit of road usage. However, this so-called "first-best" pricing method is a primarily theoretical con-

cept. In particular, the practical problems in the administration of marginal cost pricing have shifted the focus to simpler schemes of congestion taxation. So-called "second-best" alternatives, such as cordon tolls and area licensing schemes, charge mere entry to a given area, and are therefore easier to implement and less expensive to maintain (Santos and Shaffer, 2004).

Particularly restrictive and/or costly measures (e.g., city tolls, low emission zones, or operation bans for specific vehicles) show a high potential for reducing air pollution (e.g., Bigazzi and Rouleau, 2017). However, the (perceived) fairness of such measures could be limited (Eliasson, 2016; Kristoffersson et al., 2017). Also, behavioral reactions such as switching to vehicles not affected by the measure could limit the (environmental) benefits (Percoco, 2014). Overall, however, road pricing measures (in a broad sense, subsuming the aforementioned different pricing schemes) prove to be effective (Gibson and Carnovale, 2015; Santos and Shaffer, 2004; Tonne et al., 2008). Nevertheless — or due to the effectiveness in regulating road traffic — support for such costly measures, unless combined with further provisions, is rather low (Huber et al., 2020; Li and Hensher, 2012; Wicki et al., 2019a,b). Widespread support for policy measures is, however, crucial for their successful implementation (Jones, 2003; Selmoune et al., 2020).

This study analyzes which factors determine attitudes toward costly environmental measures such as city tolls. Does the reported low support permeate the entire population? Or is there a divide in the level of support and/or the preferred institutional design of tolling schemes? Using a unique combination of a factorial survey experiment and neighborhood context information, existing data can be extended to observe specific subgroups that could be differently affected by the introduction of a city road toll. In order to disentangle the underlying mechanisms explaining the amount of support for city road tolls, this analysis considers the amount of support for tolling schemes separately according to place of residence (i.e., within the hypothetical toll area vs. not) and respondents' car usage (i.e„ yes vs. no). The research question scrutinized here is therefore: Can attitudes toward a city road toll be explained primarily by general beliefs regarding policy consequences, or does individual self-interest play a substantial role in shaping attitudes?

3.2. State of research

There is a large strand of literature on attitudes toward urban road pricing, including the specific topic of city tolls. Previous research has investigated a broad set of factors ranging from determinants of acceptance for pricing schemes in general (e.g., Hensher and Li, 2013; Hysing and Isaksson, 2015; Kallbekken et al., 2013; Schade and Schlag, 2003) to attitudes toward specific schemes including changes of attitudes over time, as well as after scheme implementation (e.g., Andersson and Nässén, 2016; Börjesson et al., 2016; Börjesson and Kristoffersson, 2018; Eliasson, 2014; Hansla et al., 2017). Overall, two general schemes can be defined: flat city tolls and more differentiated charging schemes. In London, for example, a flat toll is charged for inner-city personal vehicle usage on weekdays during the day (e.g., Santos and Shaffer, 2004), whereas in Gothenburg, a time-dependent charging scheme is used (e.g., Börjesson and Kristoffersson, 2018). While the implementation of a theoretically optimal first-best pricing level requires highly differentiated city toll schemes, differentiation also increases the complexity of the scheme (cf. Link, 2015). For an overview of recent findings, see Selmoune et al. (2020). Drawing on eight cases in which city toll schemes were introduced or rejected, the authors discuss several criteria that influence the support for such systems. Among the most prominent factors driving low support are privacy concerns, lack of (perceived) fairness in the distribution of costs, uncertainty about details of the proposed scheme, and scheme complexity.

Although previous research has examined a large number of different factors affecting attitudes toward city toll schemes, some suffer from limitations that raise questions about the generalizability of the results. Many studies, for instance, are based solely on data from personal vehicle drivers (for an overview see e.g., Li and Hensher, 2012). This strong focus on motorists is reasonable when considering respondents' willingness to switch to other modes of transport. When examining attitudes toward political measures in terms of democratic opinion formation processes, however, the support of a majority of the population appears to be more relevant than the particular interests of motorists (Jaensirisak et al., 2005, p. 139). To examine the willingness to weigh different aspects (e.g., infrastructure improvements vs. additional costs) in the decision-making process, experimental techniques are needed (see e.g., Kristoffersson et al., 2017; for an overview of studies making use of experimental approaches see e.g., Li and Hensher, 2012). This is particularly

evident in Grisolía et al. (2015), who examine support for a city toll in Las Palmas de Gran Canaria, Spain. In a qualitative pre-study they find respondents to be generally opposed to the implementation of a city toll. In the choice experiment, however, they find that respondents are quite willing to seriously consider the various characteristics of a toll scheme against and (under certain conditions) evaluate the implementation of a toll positively.

Some research tackle the question of whether and to which extent differences in attitudes toward the implementation of a city toll exist within the general population. In particular, the findings of Jaensirisak et al. (2005) are noteworthy: Aside from the effect of various design features on the general support for toll systems, the authors also examine possible differences between motorists and non-motorists. Each of a total of 830 respondents from London and Leeds evaluated four experimentally-varied toll models. For each model, respondents indicated whether they would vote for its implementation, or not (i.e., yes vs. no). The authors find, as expected, stronger support for lower toll charges. In addition, non-motorists show stronger support for a city toll than do motorists. However, Jaensirisak et al. (2005) do not distinguish between respondents living within or outside of the toll area.

Studies considering place of residence relative to the toll area report mixed results. Rentziou et al. (2011) find less support for a possible city toll among residents of a potential toll area in Athens, Greece. Milenković et al. (2019), in contrast, find stronger support among residents of a possible toll area in Belgrade, Serbia. Finally, Eliasson and Jonsson (2011) find more negative evaluations of the already existing toll system among residents of the toll area in Stockholm, Sweden — but only under certain estimation model specifications. However, the authors do not examine possible differences between non-motorists and motorists living within or outside of the toll area. The present study aims to shed more light on these typically overlooked factors with a unique combination of a factorial survey experiment and geocoded neighborhood context information that can help disentangle the underlying mechanisms explaining support for city tolls.

3.3. Theoretical arguments

To better understand which aspects drive attitudes toward road pricing as a costly environmental policy, it is imperative to look closely not only at general beliefs on policy consequences (section 3.3.1), but also individual self-interest (section 3.3.2).

3.3.1. Beliefs on policy consequences

Based on the existing literature, Huber et al. (2020) identify three major influencing factors of support for policy instruments: perceived effectiveness, intrusiveness, and fairness of measures. In line with their argumentation, the presumed influences on attitudes concerning city toll schemes are briefly explained, leading to hypotheses applicable to the case study examined here.

(Perceived) Effectiveness

Huber et al. (2020) assume that individuals base their support on the extent to which a policy instrument is capable of achieving a particular goal. Based on this assumption, the authors expect that individuals who support the goal, ceteris paribus, prefer effective measures over ineffective ones. On the other hand, individuals not in support of the goal would prefer to reject the implementation of a policy than support ineffective measures (Bamberg and Rölle, 2003; Huber et al., 2020). Applied to the case of a potential city toll, individuals are expected to be more likely to support more effective toll schemes than less effective schemes. Individuals who see no need for regulation would rather oppose city tolls than support tolling schemes that achieve outcomes such as improvements in air quality or reductions to traffic volume only to a limited extent.

(Perceived) Intrusiveness

Policy interventions typically limit an individual's scope of action (Sager, 2009). The extent to which individual choices are restricted, however, varies substantially between different tolling policies. The more far-reaching the perceived consequences of a policy are in terms of, for example, limited choices or imposed costs, the less support is expected for the respective policy (Huber et al., 2020). Cherry et al. (2012) explain this relationship with social norms against (governmental) coercion. Low levels of support for coercive policies, however, could also serve to prevent oneself or others from the associated restrictions or costs (Huber et al., 2020). Concerning oneself the argument points to individuals' rational (egoistic) self-interest as a driver, whereas intentions to preserve others from imposed restrictions or costs puts the focus on underlying fairness norms. To preserve the analytical distinctiveness between different explanatory approaches, these two aspects will be discussed separately. Regarding attitudes toward a potential city toll under the intrusiveness

perspective, individuals are expected to be less likely to support toll models with more restrictive regulations compared to less restrictive schemes. Individuals who perceive the regulations of a toll model to disproportionately interfere with (not necessarily their own) personal freedom, in terms of the size of the toll area or the applicability of the toll to various vehicle types, are expected to more strongly oppose a city toll than those perceiving the regulations as less intrusive.

(Perceived) Fairness

Huber et al. (2020) further assume that individuals also consider their perception of fairness of policy instruments in their decision-making processes. Fairness can refer to procedural (e.g., opportunities to participate) as well as to distributive aspects (e.g., disproportionately affected social groups) (Preisendörfer, 2014; Tyler, 2000). If the policy instrument is evaluated as unfair, support will likely be reduced (Eriksson et al., 2006; Huber et al., 2020; Ittner et al., 2003). Individuals would therefore be more likely to support potential toll models if they perceive them as fair than models they consider unfair. Possible fairness criteria might include the spatial distribution of improvements in air quality (should there be an especially strong local improvement nearby heavily polluted crossings?) or the toll scheme pricing structure (should there be exemptions for residents?). Toll models that place a disproportionate burden on some groups are likely to be perceived as unfair, especially by the group(s) affected, and therefore receive less support than models perceived to be fair.

3.3.2. Self-interest perspective

The self-interest model proposed by Rohrschneider (1988) aims to explain political attitudes by focusing on individuals' immediate living conditions. In doing so, the model assumes rational actors interested in maximizing their own benefit. These actors form (political) attitudes depending on the extent to which their own immediate living conditions are affected, leading to the expectation that support for a political instrument increases with the extent to which it is suitable for counteracting (or at least reducing) negative impacts on respondents' living conditions. Possible factors might include environmental damage, such as noise or air pollution in the immediate area (Preisendörfer, 2017; Preisendörfer et al., 2020). The model also postulates that such damage occurring elsewhere is less of a deciding factor, as the negative

impact is not perceived to be direct.

According to the self-interest model, rational individuals consider both costs and benefits of different alternatives. The overall maximum benefit eventually determines which option individuals favor. Self-interest, integral to the formation of various attitudes, can itself be composed of various aspects. Aside from environmental damage, such as that caused by air pollution, financial damage including costs arising from political interventions to improve air quality could also be relevant to individuals' attitude formation; the introduction of additional charges for personal vehicle usage to achieve improvements in urban air quality, for example.

In order to empirically test this model for the case of potential city tolls, it is pertinent to identify subgroups whose costs and benefits resulting from the implementation of the toll scheme would differ on average, resulting in variations in self-interest and therefore attitudes concerning the toll scheme. Charges for inner-city use of motor vehicles would, for example, only result in additional costs for motorists. Thus, differentiation between motorists and non-motorists is paramount. Furthermore, one can argue that residents of the proposed toll area in particular would benefit from improvements in air quality and would therefore be more interested in measures ensuring such improvements. Crossing both dimensions (i.e., motor vehicle usage and toll area residence) defines four subgroups for which political intervention to change the status quo in the form of a city toll either lies in the individuals' self-interest, or contradicts it.

3.4. Data and methods

The presented analyses are based on a postal population survey conducted at the Department of Sociology of the LMU Munich in early summer 2018.[1] The survey was carried out using random samples of official population registers of the adult population living in Munich and several surrounding municipalities with high numbers of residents commuting to/from Munich on a regular basis. A total of 5,400 PAPI questionnaires were distributed: 3,400 to individuals living in Munich, and another 500 questionnaires each to individuals living in Gröbenzell, Landshut, Poing, and Rosenheim. This design allows for a comprehensive sample of individuals affected by a potential city toll to varying degrees, considering the issues associated

1 The survey was conducted together with Katrin Auspurg and Sabine Düval (for further details, see Thiel, 2020).

with urban mobility as well as the consequences resulting from governmental measures to improve traffic-related problems. In total, 1,335 completed questionnaires were returned, a response rate of 26% (excluding 162 undeliverable questionnaires). After further restricting the data collected due to missing values, the final analysis sample is composed of 4,654 vignette evaluations from 1,193 respondents.

A factorial survey experiment was used to measure support of a potential city toll in Munich, bringing with it advantages of an experimental design in the context of a broad population survey (e.g., Auspurg and Hinz, 2015). Each respondent evaluated four hypothetical toll models (vignettes), which differed in their respective design characteristics (dimensions). Vignettes allow the examination of the influence of various dimensions of a possible toll model on the support for the city toll. The main advantage of this approach is the independent variation of dimensions in the factorial survey experiment that are typically correlated in observational data. The dimensions considered here include a number of characteristics such as the degree of expected air quality improvements, the proposed amount of toll charges, as well as possible exemptions for residents of the toll area. These dimensions refer to the aforementioned theoretically relevant aspects of effectiveness, intrusiveness and fairness (see, section 3.3.1). Figure 3.1 depicts an example vignette as well as the 11-point response scale ranging from (−5) "definitely not" to (+5) "definitely yes", indicating the extent to which respondents support the introduction of the respective toll model.

Combination of all dimension levels results in a universe of 16,200 possible vignettes. Following the methodological recommendations for the implementation of factorial survey experiments, a sample of 1,200 different vignettes with maximum dimension level variance and the lowest possible correlations between dimensions was selected. Efficient selection increases the precision of the estimated effects of the dimensions on vignette evaluation (e.g., Auspurg and Hinz, 2015).[2] The randomized allocation of four vignettes (and, thus, of experimental stimuli) to respondents further ensures the non-correlation of vignette dimensions with individual characteristics of the respondents.

In addition to the experimentally varied dimensions of hypothetical toll models, variables distinguishing between subgroups assumed to be affected differently by the introduction of a city toll are also included. The impact of a possible tolling

2 More specifically, a D-efficient design, also orthogonalizing all two-way and three-way interactions, was implemented.

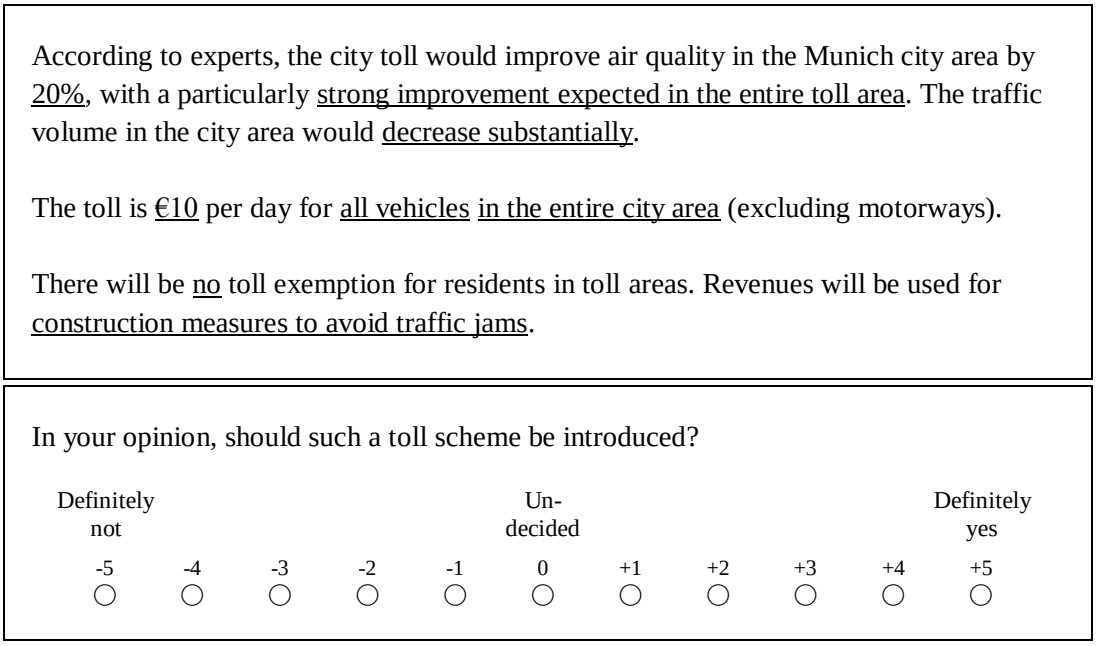

Figure 3.1.: Sample vignette with 11-point response scale

Notes: Translated version, experimentally varied dimensions underlined. For an overview of all variations, see Table 3.A1 in the appendix.

scheme must therefore be considered both in terms of residential living quality and of individual mobility behavior by encoding respondents' place of residence relative to the hypothetical toll area in the respective models (for a similar approach, see Eliasson and Jonsson, 2011). More precisely, the possible toll areas presented in the vignettes varied in size (from small to large areas) between a coverage within the historic city center, within the inner city (excluding the busy circular road "Mittlerer Ring"), within the inner city (including the "Mittlerer Ring"), and within the entire Munich city limits (excluding highways). For each of the four toll models presented, a binary variable then indicates whether the respondent lives within the described area or not. This approach allows the analyses to quantify the extent to which respondents might be affected by the proposed toll scheme at their current place of residence — particularly relevant to the self-interest model.

Individual mobility behavior was also considered. For this purpose, a binary variable indicating whether the household has access to a personal vehicle is also included (cf. Preisendörfer et al., 2020, p. 178). In particular for respondents who (at least occasionally) use a car, changes in the legal framework may have a restrictive effect on the possibilities, or the associated costs, of using that particular vehicle. As

far as choices are restricted, this also applies to non- or infrequent motorists living in the hypothetical toll area.[3]

A combination of these two binary constructs defines four subgroups, as suggested by the argumentation on individual self-interest, which take into account the varying degrees of impact the introduction of a city toll would have on the individual self-interest of the four sub-groups. Non-motorist respondents living in the toll area would benefit most from the introduction of a city toll, and are therefore expected to report the strongest support for its introduction. Motorist respondents living outside of the toll area would be most heavily affected, and are expected to show the lowest amount of support. The remaining two sub-groups are expected to report levels of support between those of the aforementioned sub-groups.

3.5. Results

This section presents the empirical results of the analysis in three steps. First, descriptive results for the entire sample as well as for the sub-groups of (non-)residents and (non-)motorists are presented. Next, the effect of specific vignette dimensions on the overall evaluation of the presented tolling schemes is analyzed. Third, an examination of whether specific dimensions affect levels of support differently among the four sub-groups. All estimates were calculated with Stata version 16 (StataCorp., 2019).

3.5.1. Descriptive results

These results cover socio-demographic characteristics of the analysis sample. The share of women is roughly 47%. On average, respondents are 50.38 years old (sd = 16.71, min. = 18, max. = 91) and the monthly household net equivalence income is on average 2,498.99 € (sd = 1377.87, min. = 400, max. = 15,000). The proportion of respondents with at least a technical college or university degree is approximately 46%, a larger share than expected in the general population of Germany. This over-representation must be taken into account when interpreting purely descriptive results; however, this artefact does not affect the robustness of the results of the

3 Further robustness checks use a variable indicating actual vehicle usage within the hypothetical toll area, as opposed to simply distinguishing whether a vehicle is present in the household. Results substantially remain robust.

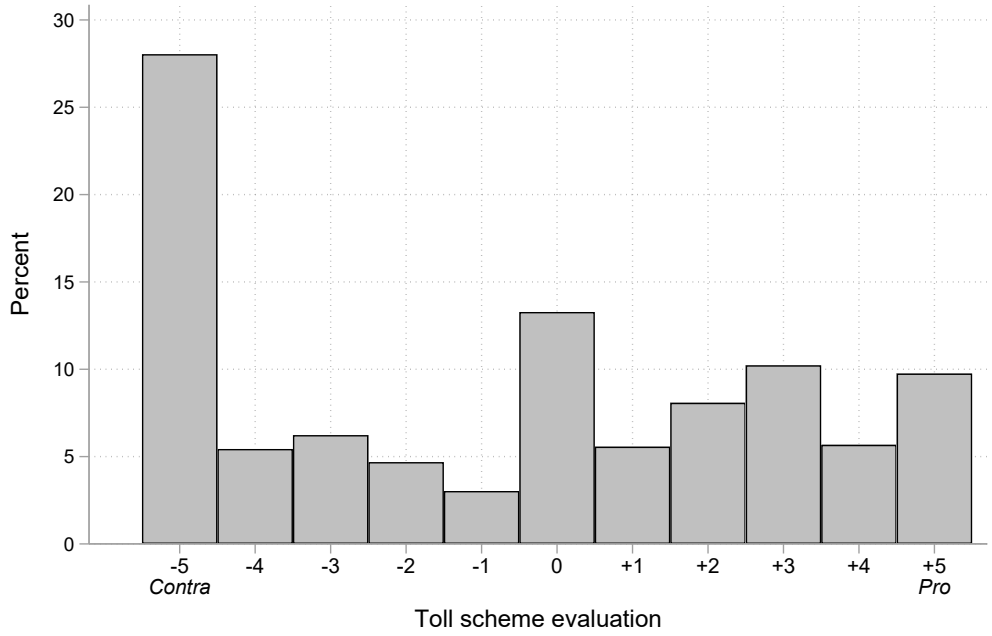

Figure 3.2.: Toll scheme evaluation distribution

Notes: Figure based on 4,654 vignette evaluations from 1,193 respondents.

factorial survey experiment.[4]

Figure 3.2 depicts the distribution of the overall evaluation of the presented toll schemes. While observed evaluations cover the entire range, overall support is rather low (mean = −0.69, sd = 3.58), mirroring findings in previous research that point to overall low support for pricing schemes. Particularly interesting, however, is the high proportion of maximum negative ratings, with 28% of all toll schemes completely rejected (−5 = "definitely not"). On the other hand, 39% of all toll schemes were evaluated (rather) positively (> 0), and another 13% without a clear tendency (0).

Figure 3.3 depicts mean evaluations of the presented toll schemes for the four sub-groups. While non-motorist respondents living outside of the hypothetical toll area (i.e., respondents not (directly) affected by a hypothetical toll scheme) evaluate the toll schemes on average with indifference (mean = −0.06, sd = 3.53), there is a clear pattern among the other three sub-groups. Motorist respondents living outside of the hypothetical toll area show the highest reluctance toward the proposed city tolls (mean = −1.00, sd = 3.55). Non-motorist respondents living within

4 For descriptive information on control variables used in further robustness checks, see Table 3.A2 in the appendix.

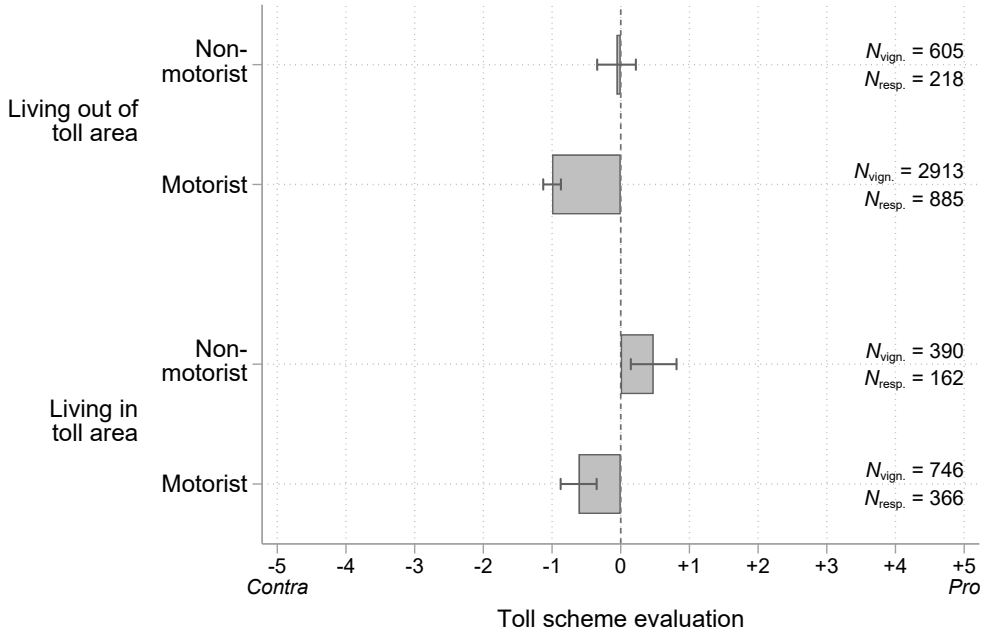

Figure 3.3.: Mean evaluation with 95% confidence intervals by sub-group

Notes: Estimates based on a total of 4,654 vignette evaluations from 1,193 respondents overall; at least 390 vignette evaluations from 162 respondents in each sub-group.

the hypothetical toll area, however, evaluate the toll schemes on average (slightly) positively (mean = 0.48, sd = 3.35) and also show a notably smaller share of maximum negative evaluations. Motorist respondents who live in the hypothetical toll area evaluate the schemes in between these two groups (mean = −0.61, sd = 3.65).

Although mean ratings of the hypothetical toll models are compared here in a merely descriptive manner, a clear pattern is still apparent. As expected from the argumentation on self-interest, respondents appear to support a city toll only to the extent that the introduction of such a toll would lie in their individual self-interest: Non-motorist respondents living in the toll area would benefit the most and represent the only sub-group that tends to support the toll on average. In contrast, motorist respondents living outside the hypothetical toll area would be most affected by restrictions, but not by improvements (e.g., air quality), and show the strongest aversion to the proposed toll schemes. The other two groups show ratings in between these two poles: Motorist respondents living in the hypothetical toll area would be affected by restrictions, but would also benefit from any improvements. This group

evaluates the toll schemes somewhat negatively on average, but not as negatively as motorists living outside the hypothetical toll area. Non-motorist respondents living outside the toll area would neither be affected by restrictions nor benefit from any improvements, reflected by their indifferent evaluations of the presented toll schemes with no clear positive or negative tendency.

3.5.2. Institutional design

To better understand what drives public support for potential city tolls, the impact of different design features on overall evaluation is estimated by regression observed ratings on all eight vignette dimensions. Random-intercept models are estimated to take the multi-level data structure into account. A likelihood ratio test shows that this significantly improves model fit when compared to the estimation of a simple linear regression (LR $\chi^2(1) = 959.00$, $p < 0.001$). As indicated by the intraclass correlation coefficient ρ, a substantial portion of about 43% of variance of the vignette evaluations can be attributed to differences between respondents. Figure 3.4 shows the results of a linear random-intercept regression model with 95% confidence intervals using cluster-robust standard errors. The results shown are based on a linear random intercept regression model without controlling for additional variables, as data stem from a factorial survey experiment randomly assigned to respondents (for exact estimates, see Table 3.A3 in the appendix). Including potential confounding factors (education, income, gender, age, children, disturbance by air pollution, environmental concern, discount rate) further restricts the analysis sample to 3,491 vignette evaluations from 889 respondents. However, the results remain substantially robust (see Table 3.A3 in the appendix).[5]

It is evident that the varied design features influence respondent ratings. Support for the toll models increases where greater improvements in air quality can be expected, whereas support decreases with higher toll charges. It is particularly revealing when these effects are related to the previously formulated theoretical expectations: The degree of expected air improvement corresponds to the effectiveness of the measure. The more effective the policy intervention (i.e., the higher the expected air quality improvement), the more strongly the toll scheme is supported. In terms of diminishing marginal utility, further improvements in air quality then

5 Results remain substantially robust when estimating regression models with respondent-specific fixed effects and standard errors clustered on both respondent and city level using the user-written Stata ado *reghdfe* (Correia, 2017).

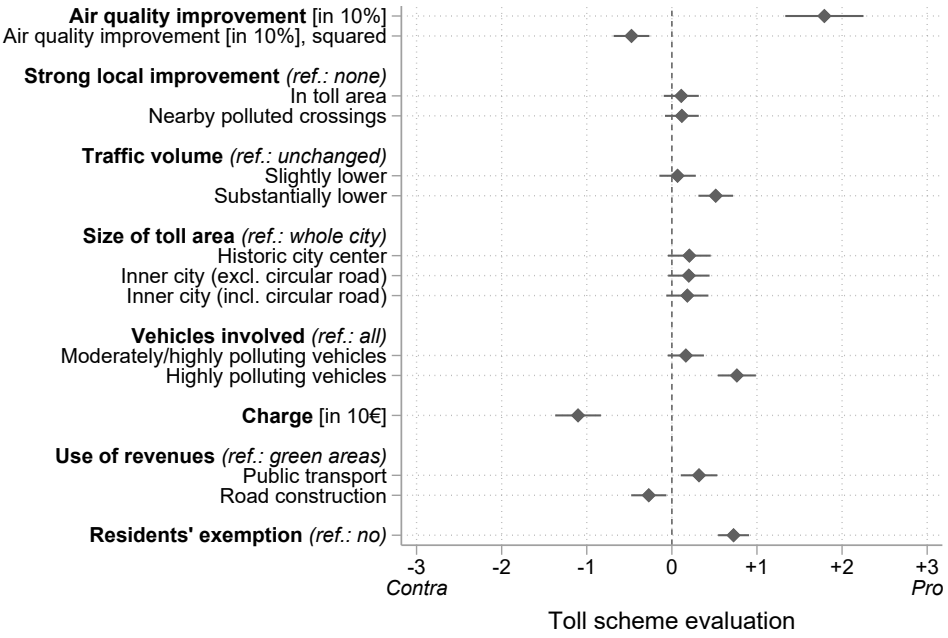

Figure 3.4.: Random-intercept model of toll scheme evaluation

Notes: Coefficients with 95% confidence intervals using cluster-robust standard errors. Coefficient plot created with the user-written Stata ado *coefplot* (Jann, 2014). Estimates based on 4,654 vignette evaluations from 1,193 respondents. The intraclass correlation coefficient is $\rho = 0.432$. For exact estimates, see Table 3.A3 in the appendix.

have a weaker, but still positive effect on the evaluation of the toll. The size of the proposed toll area can then be understood as the degree of intrusiveness of the policy intervention. A toll system that covers the entire urban area has a greater impact on previous mobility patterns than one that only extends around the historic city center. Empirically, however, there are no differences in the evaluations of the hypothetical toll schemes: Although smaller toll areas tend to be rated somewhat more positively, the effects are very small and statistically not significant ($p > 0.05$).

Where strong improvements in air quality can be expected relates to underlying fairness considerations. A substantial improvement in air quality, especially at heavily polluted intersections, can be assumed to be supported under the fairness argument in order to reduce existing disproportionate pollution. Similarly, improvements within areas where the damage caused is also paid for (i.e., within the hypothetical toll area) could be particularly welcomed. Empirical results show, however, that local improvements (at heavily polluted intersections or within the hypothetical toll area) are not supported other than a largely uniform improvement across the entire urban area. This may indicate that respondents either apply different justice principles that obscure effects, or that they place the question of the spatial distribution of improvements in air quality behind other factors influencing their evaluation.

The positive effect of residents' exemption on the support for a city toll, however, cannot be clearly attributed to the theoretical arguments. Toll models that provide an exemption for residents of the toll area are rated an average of 0.73 scale points more positively than those that do not provide an exemption ($p < 0.001$). Is this effect possibly also the result of fairness considerations, so as to not impose additional costs on residents? Or is the strong support for such an exemption rather to be understood as an expression of the individual self-interest of motorist residents who simply wish to avoid individual cost increases?

3.5.3. Subgroup differences in preferred institutional design

In order to differentiate between the two competing explanatory approaches, the analyses presented before are also estimated separately for the four sub-groups as each sub-group has been identified as having different self-interests. This approach will better distinguish which mechanism is decisive for the evaluation of the toll schemes in each sub-group.

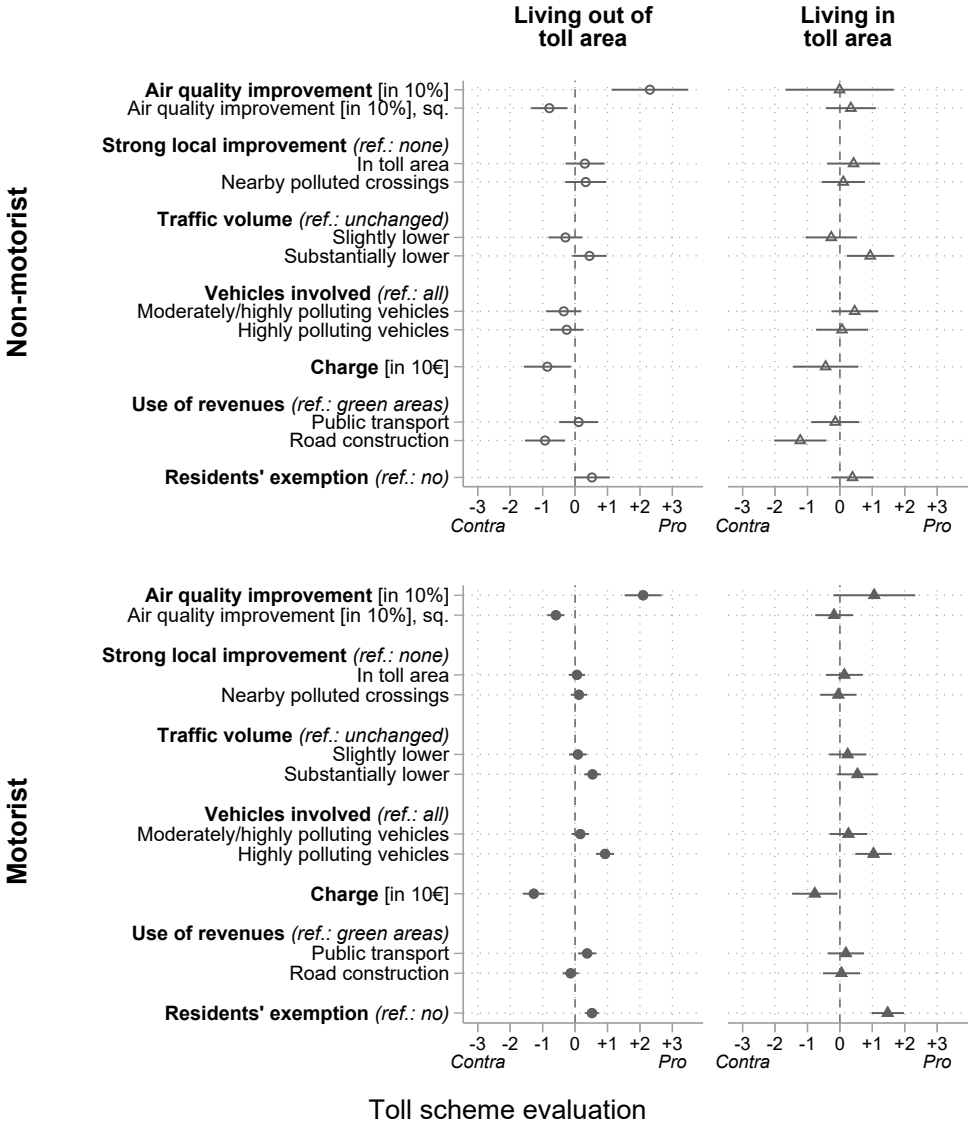

Figure 3.5.: Random-intercept models of toll scheme evaluations by sub-group

Notes: Coefficients with 95% confidence intervals using cluster-robust standard errors. Coefficient plot created with the user-written Stata ado *coefplot* (Jann, 2014). Estimates based on 4,654 vignette evaluations from 1,193 respondents total; at least 390 vignette evaluations from 162 respondents in each sub-group. Full results are shown in Table 3.A3 in the appendix.

Figure 3.5 shows the results of separate random-intercept models, regressing support for a city toll on design features of the toll scheme. Except for the exclusion of dummy variables for the respective size of the toll area (as this information is used for splitting respondents into groups of residents and non-residents), these models are estimated analogous to the previous analysis.[6] Again, coefficients with 95% confidence intervals based on cluster-robust standard errors are reported (full results are shown in Table 3.A3 in the appendix).[7]

Revealing differences in the evaluation of toll models can be observed among the four sub-groups. Motorists (living within and outside of the hypothetical toll area) are particularly in favor of toll models that apply only to very heavily polluting vehicles. In turn, non-motorist respondents are more strongly opposed to toll models that earmark revenues for road construction measures to avoid traffic jams, regardless of where they live. The exemption for residents is hardly decisive among non-motorists both inside and outside of the hypothetical toll area ($p > 0.05$). However, for motorists who live outside of the toll area, a positive, albeit relatively small, effect of offering a residents' exemption is estimated to lead to an increase in support of roughly 0.52 scale points ($p < 0.001$). For motorists living within the toll area, this effect is most pronounced, with a 1.47 scale points more positive evaluation of the toll scheme when including a residents' exemption ($p < 0.001$).

These clear differences cannot be explained only by general fairness considerations, which are assumed to prevail equally (or at least in a similar way) among respondents of all sub-groups. Rather, these patterns indicate that individual self-interest has a clear impact on the evaluation of toll schemes. The introduction of a city toll might satisfy the individual self-interest of motorists living in the toll area — provided that they themselves do not have to fear cost increases as a result. This is further expressed in very strong support for a residents' exemption, especially among motorists living in the potential toll area.

6 A binary variable is created for each of the presented toll models indicating whether the respondent lives within that same toll area.

7 Results are estimated without controlling for additional variables, as data stem from a factorial survey experiment randomly assigned to respondents. Additionally estimating separate regression models for each sub-group that also control for potential confounding factors (education, income, gender, age, children, disturbance by air pollution, environmental concern, discount rate) further restricts the analysis sample. Results remain substantially robust (see Table 3.A4 in the appendix).

3.6. Summary

The aim of the present study is to better understand which factors shape attitudes toward potential city tolls. Although the existing literature suggests overall low support for such measures, it remained unclear whether this low amount of support permeated the entire population. This investigation therefore examined whether the level of support and/or the preferred institutional design of tolling schemes differed between sub-groups of the population. Using a unique combination of a factorial survey experiment with neighborhood context information, the analysis disentangles two theoretical mechanisms: whether attitudes toward city tolls are driven mainly by general beliefs toward policy consequences, or whether individual self-interest is the driving factor. The main results are summarized in three points.

First, there is an apparent overall lack of support for city tolls in descriptive figures, although some revealing differences between sub-groups have been identified. Motorist respondents living outside of the hypothetical toll area report strong negative attitudes toward tolling schemes, whereas non-motorist respondents living inside the toll area show more positive attitudes. This suggests that respondents who personally benefit from a city toll scheme tend to support the model more than respondents being negatively affected by additional costs.

Second, institutional design is important — to some extent. In particular, the degree of improvements in air quality does substantially strengthen support for tolling schemes. In line with previous results, this finding points toward a clear preference for efficient measures. However, other beliefs on policy consequences such as the intrusiveness (e.g., size of tolling area) or fairness of the measures (e.g., distribution of air quality improvements) do not seem to carry a similar importance.

Third, which aspects of a tolling scheme are important for the support of the scheme in question differs between sub-groups. Apart from general beliefs on policy consequences, individual self-interest proves to be particularly important in the formation of attitudes toward potential city tolls. Respondents showed stronger support for the toll models associated with lower costs to them personally and/or promised substantial improvements to their personal living situation. This finding is particularly evident in the strong support for toll exemptions for residents among the group of motorist residents. However, differences between sub-groups are relatively small.

3.7. Discussion

The present study expands the current state of knowledge on support for city road tolls by several aspects. In line with previous findings, overall support for city tolls is found to be low. However, remarkable differences are found between sub-groups regarding the level of support for the measures as well as in preferred institutional design. Characteristics particularly important for support are especially driven by individual self-interest, whereas general beliefs on policy consequences shape attitudes toward city tolls only to a lesser extent. These differences in preferred characteristics could certainly be used to configure toll models capable of reaching desired goals (e.g., improvements in air quality) while also receiving widespread public support (cf. Wicki et al., 2019a).

Nevertheless, this study is not without its limitations. *First,* in contrast to simple direct questioning, the factorial survey approach allows for an experimental variation of the institutional design characteristics of a hypothetical city toll. It therefore allows to measure the effects of a broad set of dimensions on the level of support (cf. Auspurg and Hinz, 2015). However, this approach does not measure actual behavior, as no real-life consequences arise from the observed ratings. The dependent variable, support of a city toll scheme, rather reflects a behavioral intention to support the respective model. It therefore remains unclear whether the observed relationships also hold for actual support of city tolls, for example in a binding vote at the municipal level. To date, few and partly conflicting findings are available on the validity of behavioral intentions measured in survey experiments regarding actual behavior. Hainmueller et al. (2015), for example, find congruent results for behavioral intention and actual behavior, while Findley et al. (2017) report significant differences between measurements. Overall, however, for behaviors hardly associated with social desirability, behavioral intentions and actual behavior seem to match closely — at least with respect to the direction and relative strength of effects (Petzold and Wolbring, 2019, p. 8). Here, the normative expectation may be that some measure will be supported, but not specifically a city road toll. Potential social desirability issues can certainly not explain the clear differences found between sub-groups.

Second, although the 11-point response scale used here allows for fine-grained ratings, as is typical for factorial survey experiments, it is still further away from a more realistic decision situation such as a referendum vote over the introduction or

rejection of a specific toll model. Choice experiments model such a decision situation much more realistically. However, if respondents are skeptical of both alternatives in a choice experiment, they tend to choose the alternative they are less averse to, while the fact that the chosen alternative is also not preferred would be hidden by such measurement (cf. Auspurg and Liebe, 2011, p. 304). Moreover, assuming that there is no referendum (as is often the case in representative democracies), attitudes toward the implementation of a city toll remain relevant even without any actual (voting) behavior occurring.

Third, the combination of a factorial survey experiment with further information on personal vehicle usage and residential location is a major strength of the present study. This approach allows to differentiate between respondents affected to varying degrees by the advantages and disadvantages of the hypothetical city toll they are asked to rate. The presented analyses, however, are only able to differentiate superficially between motorists and non-motorists. Hence, the extent to which the intensity of (inner-city) vehicle usage might also be relevant for the evaluation of a city toll must be examined in future research. In addition, future research could make use of further context information beyond the binary differentiation of residential location relative to the hypothetical toll area. This could better exploit the potential of geocoded data, as these data would also allow, for example, the local transportation infrastructure to be considered in order to draw detailed conclusions about the availability of mobility alternatives.

Fourth, the analyses presented here do not consider possible differences in attitude formation between homeowners and renters. While the former might be interested in interventions to improve urban air quality or reduce noise exposure, thus enhancing the value of their property, the latter might be more interested in avoiding additional costs. Homeowners would then be expected to show more support for city tolls than do renters. However, as information on dwelling ownership was not obtained in this study, the analyses cannot control for such possible *home voters* (Fischel, 2001). Future studies might also consider this aspect.

Finally, the present approach does not analyze possible status quo biases in attitude formation. Previous studies have frequently found initial opposition to tolling schemes prior to implementation, followed by more positive attitudes some time afterwards (e.g., Andersson and Nässén, 2016; Börjesson et al., 2016). As a result, trial phases are a typical recommendation, offering residents and commuters the chance to familiarize themselves with the particular toll model to hopefully encourage sup-

port (Hensher and Li, 2013; Selmoune et al., 2020). However, it remains unclear to what extent the observed differences in attitudes toward city tolls might continue to exist between sub-groups. Identifying factors that persistently hinder support could shed light on aspects that, on behalf of sustainably solving traffic-related problems, should be examined more closely. Therefore, future research could utilize the unique combination of a factorial survey experiment and (geocoded) neighborhood context information presented here within a longitudinal framework to further investigate possible changes in attitudes.

References

Andersson, David and Jonas Nässén. 2016. "The Gothenburg congestion charge scheme: A pre–post analysis of commuting behavior and travel satisfaction." *Journal of Transport Geography* 52:82–89.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial survey experiments*. Thousand Oaks, California: Sage.

Auspurg, Katrin and Ulf Liebe. 2011. "Choice-Experimente und die Messung von Handlungsentscheidungen in der Soziologie." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 63:301–314.

Bamberg, Sebastian and Daniel Rölle. 2003. "Determinants of people's acceptability of pricing measures: Replication and extension of a causal model", pp. 235–248. In: Schade, Jens and Bernhard Schlag (eds.): *Acceptability of Transport Pricing Strategies*. Oxford: Elsevier.

Bigazzi, Alexander Y. and Mathieu Rouleau. 2017. "Can traffic management strategies improve urban air quality? A review of the evidence." *Journal of Transport & Health* 7:111–124.

Börjesson, Maria, Jonas Eliasson, and Carl Hamilton. 2016. "Why experience changes attitudes to congestion pricing: The case of Gothenburg." *Transportation Research Part A: Policy and Practice* 85:1–16.

Börjesson, Maria and Ida Kristoffersson. 2018. "The Swedish congestion charges: Ten years on." *Transportation Research Part A: Policy and Practice* 107:35–51.

Cherry, Todd L., Steffen Kallbekken, and Stephan Kroll. 2012. "The acceptability of efficiency-enhancing environmental taxes, subsidies and regulation: An experimental investigation." *Environmental Science & Policy* 16:90–96.

Correia, Sergio. 2017. "Linear models with high-dimensional fixed effects: An efficient and feasible estimator." Manuscript. `http://scorreia.com/research/hdfe.pdf`.

Eliasson, Jonas. 2014. "The role of attitude structures, direct experience and reframing for the success of congestion pricing." *Transportation Research Part A: Policy and Practice* 67:81–95.

Eliasson, Jonas. 2016. "Is congestion pricing fair? Consumer and citizen perspectives on equity effects." *Transport Policy* 52:1–15.

Eliasson, Jonas and Lina Jonsson. 2011. "The unexpected "yes": Explanatory factors behind the positive attitudes to congestion charges in Stockholm." *Transport Policy* 18:636–647.

Ellison, Richard B., Stephen P. Greaves, and David A. Hensher. 2013. "Five years of London's low emission zone: Effects on vehicle fleet composition and air quality." *Transportation Research Part D: Transport and Environment* 23:25–33.

Eriksson, Louise, Jörgen Garvill, and Annika M. Nordlund. 2006. "Acceptability of travel demand management measures: The importance of problem awareness, personal norm, freedom, and fairness." *Journal of Environmental Psychology* 26:15–26.

Fensterer, Veronika, Helmut Kuchenhoff, Verena Maier, Heinz-Erich Wichmann, Susanne Breitner, Annette Peters, Jianwei Gu, and Josef Cyrys. 2014. "Evaluation of the impact of low emission zone and heavy traffic ban in Munich (Germany) on the reduction of PM_{10} in ambient air." *International Journal of Environmental Research and Public Health* 11:5094–5112.

Findley, Michael G., Brock Laney, Daniel L. Nielson, and Jason C. Sharman. 2017. "External validity in parallel global field and survey experiments on anonymous incorporation." *The Journal of Politics* 79:856–872.

Fischel, William A. 2001. *The homevoter hypothesis: How home values influence local government taxation, school finance, and land-use policies*. Cambridge, MA: Harvard University Press.

Franzen, Axel and Dominikus Vogl. 2013. "Two decades of measuring environmental attitudes: A comparative analysis of 33 countries." *Global Environmental Change* 23:1001–1008.

Gibson, Matthew and Maria Carnovale. 2015. "The effects of road pricing on driver behavior and air pollution." *Journal of Urban Economics* 89:62–73.

Grisolía, José M., Francisco López, and Juan de Dios Ortúzar. 2015. "Increasing the acceptability of a congestion charging scheme." *Transport Policy* 39:37–47.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating vignette and conjoint survey experiments against real-world behavior." *Proceedings of the National Academy of Sciences* 112:2395.

Hansla, André, Erik Hysing, Andreas Nilsson, and Johan Martinsson. 2017. "Explaining voting behavior in the Gothenburg congestion tax referendum." *Transport Policy* 53:98–106.

Hardin, Garret. 1968. "The tragedy of the commons." *Science* 162:1243–1248.

Hensher, David A. and Zheng Li. 2013. "Referendum voting in road pricing reform: A review of the evidence." *Transport Policy* 25:186–197.

Holman, Claire, Roy Harrison, and Xavier Querol. 2015. "Review of the efficacy of low emission zones to improve urban air quality in European cities." *Atmospheric Environment* 111:161–169.

Huber, Robert A., Michael L. Wicki, and Thomas Bernauer. 2020. "Public support for environmental policy depends on beliefs concerning effectiveness, intrusiveness, and fairness." *Environmental Politics* 29:649–673.

Hysing, Erik and Karolina Isaksson. 2015. "Building acceptance for congestion charges – the Swedish experiences compared." *Journal of Transport Geography* 49:52–60.

Ittner, Heidi, Ralf Becker, and Elisabeth Kals. 2003. "Willingness to support traffic policy measures: The role of justice", pp. 249–265. In: Schade, Jens and Bernhard Schlag (eds.): *Acceptability of Transport Pricing Strategies*. Oxford: Elsevier.

Jaensirisak, Sittha, Mark Wardman, and Anthony D. May. 2005. "Explaining variations in public acceptability of road pricing schemes." *Journal of Transport Economics and Policy* 39:127–154.

Jann, Ben. 2014. "Plotting regression coefficients and other estimates." *The Stata Journal* 14:708–737.

Jones, Peter M. 2003. "Acceptability of road user charging: meeting the challenge", pp. 27–65. In: Schade, Jens and Bernhard Schlag (eds.): *Acceptability of Transport Pricing Strategies*. Oxford: Elsevier.

Kallbekken, Steffen, Jorge H. García, and Kristine Korneliussen. 2013. "Determinants of public support for transport taxes." *Transportation Research Part A: Policy and Practice* 58:67–78.

Kristoffersson, Ida, Leonid Engelson, and Maria Börjesson. 2017. "Efficiency vs equity: Conflicting objectives of congestion charges." *Transport Policy* 60:99–107.

Lee, Shin. 2018. "Transport policies, induced traffic and their influence on vehicle emissions in developed and developing countries." *Energy Policy* 121:264–274.

Li, Zheng and David A. Hensher. 2012. "Congestion charging and car use: A review of stated preference and opinion studies and market monitoring evidence." *Transport Policy* 20:47–61.

Link, Heike. 2015. "Is car drivers' response to congestion charging schemes based on the correct perception of price signals?" *Transportation Research Part A: Policy and Practice* 71:96–109.

Milenković, Marina, Draženko Glavić, and Milica Maričić. 2019. "Determining factors affecting congestion pricing acceptability." *Transport Policy* 82:58–74.

Morfeld, Peter, David A. Groneberg, and Michael F. Spallek. 2014a. "Effectiveness of low emission zones: Large scale analysis of changes in environmental NO_2, NO and NO_x concentrations in 17 German cities." *PLOS ONE* 9:e102999.

Morfeld, Peter, David A. Groneberg, and Michael F. Spallek. 2014b. "Effectiveness of low emission zones of stage 1: Analysis of the changes in fine dust concentrations (PM_{10}) in 19 German cities." *Pneumologie* 68:173–186.

Morfeld, Peter, David A. Groneberg, and Michael F. Spallek. 2015. "Letter to the Editor: On the effectiveness of low emission zones." *Atmospheric Environment* 122:569–570.

Percoco, Marco. 2014. "The effect of road pricing on traffic composition: Evidence from a natural experiment in Milan, Italy." *Transport Policy* 31:55–60.

Petzold, Knut and Tobias Wolbring. 2019. "What can we learn from factorial surveys about human behavior? A validation study comparing field and survey experiments on discrimination." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 15:19–30.

Pigou, Arthur C. 1929. *The economics of welfare.* London: Macmillan.

Pope, C. Arden and Douglas W. Dockery. 2006. "Health effects of fine particulate air pollution: Lines that connect." *Journal of Air & Waste Management Association* 56:709–742.

Preisendörfer, Peter. 2014. "Umweltgerechtigkeit: Von sozial-räumlicher Ungleichheit hin zu postulierter Ungerechtigkeit lokaler Umweltbelastungen." *Soziale Welt* 65:25–45.

Preisendörfer, Peter. 2017. *Personal exposure to unfavorable environmental conditions: Does it stimulate environmental activism?* Berlin/Boston: De Gruyter Oldenbourg.

Preisendörfer, Peter, Lucie Herold, and Karin Kurz. 2020. "Road traffic and aircraft noise as drivers of environmental protest?" *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72:165–191.

Qadir, Raeed M., Gülcin Abbaszade, Jürgen Schnelle-Kreis, Judith C. Chow, and Ralf Zimmermann. 2013. "Concentrations and source contributions of particulate organic matter before and after implementation of a low emission zone in Munich, Germany." *Environmental Pollution* 175:158–167.

Rentziou, Aikaterini, Christina Milioti, Konstantina Gkritza, and Matthew G. Karlaftis. 2011. "Urban road pricing: Modeling public acceptance." *Journal of Urban Planning and Development* 137:56–64.

Rohrschneider, Robert. 1988. "Citizens' attitudes toward environmental issues: Selfish or selfless?" *Comparative Political Studies* 21:347–367.

Sager, Fritz. 2009. "Governance and coercion." *Political Studies* 57:537–558.

Santos, Georgina and Blake Shaffer. 2004. "Preliminary results of the London congestion charging scheme." *Public Works Management & Policy* 9:164–181.

Schade, Jens and Bernhard Schlag. 2003. "Acceptability of urban transport pricing strategies." *Transportation Research Part F: Traffic Psychology and Behaviour* 6:45–61.

Selmoune, Aya, Qixiu Cheng, Lumeng Wang, and Zhiyuan Liu. 2020. "Influencing factors in congestion pricing acceptability: A literature review." *Journal of Advanced Transportation* 2020:1–11.

StataCorp. 2019. *Stata statistical software: Release 16*. College Station, TX: StataCorp LP.

Thiel, Fabian. 2020. "Die Low-Cost-Hypothese. Ein empirischer Test am Beispiel der Befürwortung einer City-Maut." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72:429–453.

Tonne, Cathryn, Sean Beevers, Ben Armstrong, Frank Kelly, and Paul Wilkinson. 2008. "Air pollution and mortality benefits of the London congestion charge: Spatial and socioeconomic inequalities." *Occupational and Environmental Medicine* 65:620–627.

Tyler, Tom R. 2000. "Social justice: Outcome and procedure." *International Journal of Psychology* 35:117–125.

Walters, Alan A. 1961. "The theory and measurement of private and social cost of highway congestion." *Econometrica* 29:676–699.

Wicki, Michael L., Lukas Fesenfeld, and Thomas Bernauer. 2019a. "In search of politically feasible policy-packages for sustainable passenger transport: Insights from choice experiments in China, Germany, and the USA." *Environmental Research Letters* 14:084048.

Wicki, Michael L., Robert A. Huber, and Thomas Bernauer. 2019b. "Can policy-packaging increase public support for costly policies? Insights from a choice experiment on policies against vehicle emissions." *Journal of Public Policy* 40:599–625.

3.A. Appendix

Table 3.A1.: Overview of dimensions and levels experimentally varied in vignette descriptions

	Dimension	Level	Σ
1	Air quality improvements in entire city area[a]	• By 1% • [By 5%] • By 10% • [By 15%] • By 20%	5
2	Locations with large improvements	• In toll area • Nearby polluted crossings • None (even across entire city area)	3
3	Traffic volume	• Unchanged • Slightly lower • Substantially lower	3
4	Size of toll area	• Historic city center (incl. "Altstadtring") • Inner city (excl. "Mittlerer Ring") • Inner city (incl. "Mittlerer Ring") • Whole city area (excl. highways)	4
5	Vehicles involved	• All • Moderately and highly polluting vehicles (all without "blue badge") • Highly polluting vehicles (all without "green badge")	3
6	Charge per day[a]	• 1 € • [3 €] • 5 € • [7 €] • 10 €	5
7	Use of revenue	• Investments in public transport • Investments in green areas in the city • Road construction work to reduce congestion	3
8	Residents' exemption	• Yes • No	2
Σ	Vignette universe ($5 \times 3 \times 3 \times 4 \times 3 \times 5 \times 3 \times 2$)		16,200

Notes: Translated version. [a]In order to better estimate possible non-linear effects, the degree of improvement in air quality in the city area and the toll rate were varied between three and five possible values (values in square brackets) for half of the respondents.

Table 3.A2.: Descriptive statistics

	Mean	SD	Min.	Max.	Description
Gender	0.47		0	1	0=“male”, 1=“female”
Age	50.38	16.71	18	91	In years
Education	0.46		0	1	At least technical college or university degree (0=“no”, 1=“yes”)
Income	2498.99	1377.87	400	15000	Household net equivalence income in euro
Children	0.22		0	1	0=“no”, 1=“yes”
Disturbance by air pollution	2.24	1.25	1	5	1=“not at all disturbed”, 5=“strongly disturbed”
Environmental concern	3.78	0.74	1	5	1=“low”, 5=“high”
Discount rate	0.48	0.42	0.025	1	Subjective discount rate

Notes: Descriptive statistics based on 889 respondents.

Table 3.A3.: Random-intercept models of support for city tolls, estimated for the entire sample and separately by sub-group

	(1)	(2)	(3)	(4)	(5)	(6)
	Base model	With controls	By subgroup			
			Living out of toll area; Non-motorist	Living in toll area; Non-motorist	Living out of toll area; Motorist	Living in toll area; Motorist
Air quality improvement [in 10%]	1.792***	1.828***	2.307***	−0.010	2.100***	1.058
	(0.235)	(0.273)	(0.602)	(0.851)	(0.291)	(0.644)
Air quality improvement [in 10%], sq.	−0.474***	−0.456***	−0.792**	0.337	−0.590***	−0.179
	(0.107)	(0.124)	(0.286)	(0.393)	(0.133)	(0.297)
Strong local improvement (ref.: none)						
In toll area	0.111	0.031	0.302	0.423	0.057	0.138
	(0.105)	(0.124)	(0.304)	(0.417)	(0.129)	(0.290)
Nearby polluted crossings	0.118	0.029	0.329	0.106	0.119	−0.047
	(0.101)	(0.119)	(0.322)	(0.340)	(0.128)	(0.285)
Traffic volume (ref.: unchanged)						
Slightly lower	0.067	0.183	−0.295	−0.265	0.088	0.241
	(0.109)	(0.127)	(0.268)	(0.405)	(0.140)	(0.292)
Substantially lower	0.515***	0.561***	0.446+	0.936*	0.539***	0.542+
	(0.104)	(0.118)	(0.270)	(0.369)	(0.131)	(0.318)
Size of toll area (ref.: whole city)						
Historic city center	0.205	0.206				
	(0.128)	(0.150)				
Inner city (excl. “Mittlerer Ring”)	0.200	0.182				
	(0.124)	(0.144)				

(continued on next page)

(continued)

	(1) Base model	(2) With controls	(3) By sub-group	(4)	(5)	(6)
			Living out of toll area; Non-motorist	Living in toll area; Non-motorist	Living out of toll area; Motorist	Living in toll area; Motorist
Inner city (incl. "Mittlerer Ring")	0.182	0.287^{+}				
	(0.126)	(0.148)				
Vehicles involved (ref.: all)						
Moderately/highly polluting vehicles	0.165	0.053	−0.348	0.456	0.163	0.265
	(0.108)	(0.127)	(0.277)	(0.366)	(0.136)	(0.297)
Highly polluting vehicles	0.764^{***}	0.710^{***}	−0.257	0.066	0.926^{***}	1.037^{***}
	(0.114)	(0.134)	(0.265)	(0.408)	(0.142)	(0.288)
Charge [in 10€]	-1.102^{***}	-1.098^{***}	-0.854^{*}	−0.439	-1.272^{***}	-0.773^{*}
	(0.137)	(0.159)	(0.369)	(0.517)	(0.169)	(0.357)
Use of revenue (ref.: green areas)						
Public transport	0.319^{**}	0.364^{**}	0.111	−0.142	0.373^{**}	0.185
	(0.109)	(0.126)	(0.307)	(0.378)	(0.141)	(0.285)
Road construction	-0.271^{*}	-0.229^{+}	-0.926^{**}	-1.221^{**}	−0.137	0.049
	(0.105)	(0.121)	(0.316)	(0.411)	(0.125)	(0.291)
Residents' exemption (ref.: no)	0.725^{***}	0.775^{***}	0.521^{+}	0.388	0.522^{***}	1.474^{***}
	(0.093)	(0.108)	(0.282)	(0.332)	(0.113)	(0.254)
Constant	-2.274^{***}	-6.530^{***}	−0.653	−0.272	-2.486^{***}	-2.548^{***}
	(0.196)	(0.550)	(0.525)	(0.626)	(0.227)	(0.520)
Additional controls	No	Yes	No	No	No	No

(continued on next page)

(continued)

	(1) Base model	(2) With controls	(3) By sub-group	(4)	(5)	(6)
			Living out of toll area; Non-motorist	Living in toll area; Non-motorist	Living out of toll area; Motorist	Living in toll area; Motorist
ρ	0.432	0.358	0.417	0.418	0.450	0.381
Overall R^2	0.050	0.142	0.057	0.051	0.063	0.062
N (evaluations)	4654	3491	605	390	2913	746
N (respondents)	1193	889	218	162	885	366

Notes: Coefficients and cluster-robust standard errors in parentheses. Column (1) reports results of the base model. Column (2) reports results including potential confounding factors (education, income, gender, age, children, disturbance by air pollution, environmental concern, discount rate). Columns (3) to (6) report results of the base model estimated separately for subgroups. ρ represents the intraclass correlation coefficient.

$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.A4.: Random-intercept models of support for city tolls, estimated separately for sub-groups

	(7)	(8)	(9)	(10)
	Living out of toll area; Non-motorist	Living in toll area; Non-motorist	Living out of toll area; Motorist	Living in toll area; Motorist
Air quality improvement	2.800***	−0.199	2.221***	0.539
[in 10%]	(0.743)	(0.986)	(0.341)	(0.715)
Air quality improvement	−1.012**	0.399	−0.611***	0.139
[in 10%], sq.	(0.361)	(0.452)	(0.155)	(0.325)
Strong local improvement (ref.: none)				
In toll area	0.120	0.169	−0.004	0.180
	(0.364)	(0.472)	(0.151)	(0.330)
Nearby polluted crossings	−0.036	0.055	0.084	−0.166
	(0.371)	(0.424)	(0.150)	(0.326)
Traffic volume (ref.: unchanged)				
Slightly lower	−0.012	−0.157	0.157	0.410
	(0.338)	(0.481)	(0.160)	(0.329)
Substantially lower	0.453	0.746	0.556***	0.765*
	(0.328)	(0.454)	(0.149)	(0.349)
Vehicles involved (ref.: all)				
Moderately/highly polluting vehicles	−0.725*	0.593	0.026	0.108
	(0.334)	(0.425)	(0.158)	(0.328)
Highly polluting vehicles	−0.144	−0.066	0.804***	1.020**
	(0.368)	(0.483)	(0.165)	(0.327)
Charge [in 10€]	−1.166**	−0.453	−1.288***	−0.550
	(0.428)	(0.588)	(0.197)	(0.386)
Use of revenue (ref.: green areas)				
Public transport	0.015	0.065	0.457**	0.113
	(0.380)	(0.449)	(0.161)	(0.295)
Road construction	−0.656+	−1.004*	−0.154	0.124
	(0.376)	(0.492)	(0.143)	(0.332)
Residents' exemption	0.646+	0.245	0.594***	1.279***
(ref.: no)	(0.348)	(0.396)	(0.132)	(0.294)
Constant	−5.693***	−2.044	−6.440***	−8.387***
	(1.639)	(1.926)	(0.698)	(1.092)
Additional controls	Yes	Yes	Yes	Yes
ρ	0.296	0.351	0.402	0.238
Overall R^2	0.194	0.103	0.131	0.220
N (evaluations)	426	299	2188	578
N (respondents)	154	121	667	280

Notes: Coefficients and cluster-robust standard errors in parentheses. Models also include potential confounding factors (education, income, gender, age, children, disturbance by air pollution, environmental concern, discount rate). ρ represents the intraclass correlation coefficient.
$^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

4. Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination

Dieses Kapitel ist erschienen als:

Auspurg, Katrin, Andreas Schneck, und Fabian Thiel. 2020. „Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination.“ *Research in Social Stratification and Mobility* 65:100444. DOI: `https://doi.org/10.1016/j.rssm.2019.100444`

Zusammenfassung

In diesem Beitrag wird ein möglicher Samplingbias bei Feldexperimenten zu ethnischer Diskriminierung auf dem Wohnungsmarkt durch eine einzigartige Kombination aus einem groß angelegten Feldexperiment auf dem deutschen Mietwohnungsmarkt ($N = 2,992$ getestete Wohnungen) und Daten des Internet-Wohnungsmarkts, aus dem die Wohnungen gesampelt wurden (Beobachtung der gesamten Plattform über ein Jahr), untersucht. Bisher wurden in den meisten Feldexperimenten Stichproben auf der Ebene der Anbieter (nicht der Wohnungen) gezogen und die getesteten Anbieter innerhalb eines kurzen Feldzeitraums ausgewählt (Punkt-Sampling). Dies führte wahrscheinlich zu einer Überrepräsentation von kleinen (privaten) Anbietern und Wohnungen, die bereits seit längerer Zeit inseriert waren, sowie zu einer Unterrepräsentation von großen Anbietern und neuen Leerständen. Wir analysieren, wie sich beide Aspekte auf die Messung der ethnischen Diskriminierung und die ihr zugrunde liegenden Mechanismen auswirken. Bei unserer Fallstudie zum deutschen Wohnungsmarkt beobachteten wir zunächst die erwartete Stichprobenverzerrung: Es gab eine starke Übererfassung von kleinen (privaten) Anbietern und von Wohnungen, die bereits seit langem angeboten wurden. Zweitens wurde festgestellt, dass diese Verzerrung nur geringe Auswirkungen auf das deskriptive Ergebnis einer erheblichen Diskriminierung türkischer Bewerber hat. Es gab nur eine leichte Tendenz, dass die Diskriminierung bei kleinen Anbietern und bei erst relativ kurz ausgeschriebenen Angeboten (d.h. bei neuen Angeboten) höher war. Insgesamt kommen wir zu dem Schluss, dass die Erkenntnisse über ethnische Diskriminierung und die ihr zugrunde liegenden Risikofaktoren bemerkenswert robust gegenüber der verwendeten Stichprobentechnik sind. Obwohl es keine Anzeichen für eine schwerwiegende Verzerrung gab, zeigen unsere Marktdaten Möglichkeiten auf, die Auswirkungen von Marktbedingungen zu testen, die andernfalls oft unbemerkt bleiben.

Abstract

This paper explores a possible sampling bias in field experiments through a unique combination of a large-scale field experiment on ethnic discrimination in the German rental housing market ($N = 2,992$ tested apartments) and data on the internet housing market where the apartments were sampled from (observation of the whole platform for about one year). Up to now, most field experiments sampled on the level of suppliers (not apartments) and selected the tested suppliers within a short field period (point sampling). This probably led to an over-representation of small suppliers and apartments that had already been advertised for a long time, resulting in an under-representation of large suppliers and new vacancies. We analyze how both issues affect the measurement of ethnic discrimination and its underlying mechanisms. With our case study on the German housing market, we first observed the expected sampling bias: There was a strong over-sampling of small suppliers and apartments already offered for a long time. Second, this bias was found to have only little impact on the descriptive result of substantial discrimination against Turkish applicants. There was only a slight tendency that discrimination was higher for small suppliers and offers advertised for a relatively short time (i.e. new vacancies). Overall, we conclude that evidence on ethnic discrimination and its underlying risk factors is remarkably robust to the used sampling technique. Although there were no indications of a severe bias, our market data illustrate opportunities to test the effects of market conditions that often remained unnoticed.

4.1. Introduction

This study presents a unique combination of a field experiment on ethnic discrimination with big data on the internet platform where the tested housing units were sampled from. These data inform on the size of suppliers and the time market offers were advertised online: Not only for the sample used in the field experiments, but on all offers advertised on the markt platform. This combination allows us to explore whether standard sampling techniques used in field experiments lower their external validity. As we will argue, these techniques frequently over-sampled small suppliers. When using short field periods, they additionally over-sampled offers with a low success rate (i.e. long advertisement time). This over-representation of units with a long survival time has been termed *length bias* in the statistical literature on observational data (van Es et al., 2000). To our knowledge, there is so far no literature that explores these sampling issues for field experiments. We will illustrate the resulting biases with a case study on the German rental housing market. However, the main results probably also generalize to other markets, such as the labor or product market, as the sampling strategies tested in our study are standard to field experimental methods in general.

Regarding housing markets, there is large evidence for Western countries that migrants are disadvantaged compared to the majority population. They live, for instance, in apartments with relatively few amenities; they face larger housing costs; and they are also more likely to live in poorer neighborhoods than the majority population (for statistics on Germany: Drever and Clark, 2002; Harrison et al., 2005). These disparities intersect with other social inequalities, such as poor health, education, and labor market outcomes (Galster, 1992, 1996; Pager, 2008; Turner et al., 2002). Given that, and given an increasing shortage of affordable housing in many Western metropolises with booming labor markets (for Europe: Ball, 2016; for the U.S.: Metcalf, 2018), access to housing is increasingly seen as a crucial factor for determining ethnic minorities' position in the social stratification.

It is thus not surprising that there is substantial interest in the question of the extent to which access to housing is restricted by ethnic discrimination. Starting in the 1960s, there is a large strand of research that used field experiments to answer this question (for two recent meta-analyses: Auspurg et al., 2019b; Flage, 2018). Following the standard procedure in experiments, most scholars focused solely on the internal validity (randomization) of the experiments and ignored how

non-representativeness of their selected (housing) units could have impacted their findings. With the possibility of running the field experiments in natural environments, the experiments were simply thought to allow for a sufficiently high amount of external validity. However, when offers of suppliers with a different tendency to discriminate have a distinct probability of being sampled for the experiment, the external validity might be threatened (Bell and Stuart, 2016). In this article, we will discuss in particular two threats to the validity, which are—to the best of our knowledge—missing in the literature on field experiments.

First, while some researchers sampled each housing unit with an equal probability (never minding whether they belong to the same or different suppliers), other researchers sampled each supplier only once. The latter is mainly done for ethical reasons: Sampling each supplier only once means lowering the loss of time and other burdens for the landlords and agencies that are tested in the experiments (for a general discussion of ethical issues in field experiments: Riach and Rich, 2004). However, sampling on the level of suppliers (and not apartments) might also impact the external validity. Compared to the population of all housing vacancies, this will result in an over-representation of vacancies that are offered by small suppliers.

Second, sampling always took place during a limited field period that lasted between one week and several months. In this period, a cross-section of units advertised on the market was sampled. Using such a cross-section induces what has been called a *length bias* in the statistical literature on sampling (van Es et al., 2000): The probability of being sampled is proportional to the length of time a housing unit stays on the market (is advertised). Thus, units with a low success rate (long survival time on the market platform) are more likely to be sampled than units with a high success rate (low survival time).

Given both issues, the tested units certainly do not represent a random selection from the population of (housing) offers. Small suppliers are mostly private landlords that follow less formal standards than commercial agencies, while units with a low success rate are likely those that are of relatively low quality (given the price) and/or located in deprived areas where it is difficult to find any renters. As we will argue in more detail later on, there are good reasons to believe that this translates to a biased measurement of discrimination. If this is true, the inconsistency of findings on the amount and kind of discrimination (tastes or statistical discrimination) might partly stem from different sampling techniques.

To test our assumptions, we used a novel combination of field experimental data

and market data. In 2015, we conducted a large-scale field experiment (e-mail correspondence test) on ethnic discrimination in the rental housing market in Germany. What is specific to our study is that we combined this experiment with rich data on the internet platform where the apartments were advertised. For about one year, we observed the whole internet platform with more than one million advertised housing units. In this study, we analyze 2,992 units tested in Western Germany where we have full information on the size of the supplier (measured by the number of offers advertised during our one-year observation period) and length of the advertisement time before experimental treatment (i.e. we sent our e-mails to apply for the apartment, using Turkish and German identities). To what extent is the sample of our field experiment biased regarding the size of suppliers and/or the length of advertisements? Does this affect the observed discrimination rates? And is there evidence that sampling bias affects not only the level but also the observed mechanisms underlying discrimination, such as the incidence of statistical or taste-based discrimination?

4.2. State of the Art: Sampling Strategies Used in Field Experiments

Field experiments were developed as a particularly appealing method to measure discrimination in (housing) markets (Pager, 2008).[1] In prior decades, researchers mainly used in-person audits (e.g. Yinger, 1986), where test persons with different ethnic backgrounds apply to the same housing units. With apartments becoming increasingly advertised on the internet, audits have been more and more replaced by correspondence studies where standardized, written e-mail applications are sent out with the names of applicants signaling their ethnicity (for a review of different field experimental methods see e.g. Bertrand and Duflo, 2017). In these experiments, discrimination is typically measured by the differences in the response probabilities to the e-mails sent by minority versus majority applicants. Due to the experimental manipulation of ethnicity, these e-mail experiments provide a high amount of confidence (internal validity) that the ethnicity did cause the variation in the observed

1 Observational data suffer from strong problems of unobserved heterogeneity: Ethnicity is correlated with many (unobserved) factors such as monetary resources, social networks and willingness to pay for housing (for an overview on different methods to measure discrimination see e.g. Pager and Shepherd, 2008).

outcomes.

Meta-analyses of these experiments document a substantial amount of discrimination for the U.S. and Europe, with a high amount of variation between but also within countries (Auspurg et al., 2019b; Flage, 2018). Audit studies are seen to suffer from methodological weaknesses that may threaten the internal validity of the results (such as experimenter demand effects and unobserved treatment heterogeneity, see e.g. Heckman, 1998). In the following, we therefore review only the most recent e-mail correspondence tests published between 2010 and 2019 (see Table 4.1).

In this period, $N = 20$ e-mail correspondence tests covered a wide range of European countries and the U.S., as well as a wide range of tested ethnicities. Overall, this specific subpopulation of more recent tests provided very robust evidence for ethnic discrimination. Compared to the response probability of the majority applicant, the response probability for the minority applicant was on average 11 percentage points lower. As can be seen from Table 4.1, the studies varied in the tested ethnicities, the year of study, and also the experimental design: Whereas in some studies only one application was sent per vacancy (between-design), in other studies various applications with different ethnicities were sent (within-design). Within-designs offer greater statistical power to detect discrimination (Charness et al., 2012, p. 2). A downside is, however, the larger probability of being detected by suppliers as they receive at least two similar requests.

The pros and cons of these different experimental designs and also differences across ethnicities and time trends have already been discussed in the literature (e.g. Heckman, 1998; Vuolo et al., 2016). We are, however, not aware of any studies focusing on the sampling techniques that are also summarized in Table 4.1. First, there is variation in the used sampling frame (sampling on the level of suppliers or apartments): Whereas in about half of the studies the researchers sent only one application to each supplier—although she or he might have advertised more apartments—(*supplier-sampling*), in the other half of studies the researchers sampled every apartment with an equal probability (*apartment-sampling*).

Second, the length of the field period when apartments were sampled also differed across the studies, ranging from one week (Mazziotta et al., 2015) to more than half a year (Bosch et al., 2015). As we will demonstrate later on in more detail, when using very short sampling periods, in particular apartments with a high success rate have already disappeared from the market platform when the sampling period sets in. That is, new and short vacancies will be under-sampled, while apartments

with a long advertisement time will be over-sampled. If the supplier-sampling or length bias have a meaningful impact on the observed level of discrimination, part of the cross-study variation shown in Table 4.1 would be artificial, simply caused by different sampling techniques.

4.3. Theoretical Background: Do Sampling Techniques Limit External Validity?

Researchers frequently make very general conclusions on the incidence of discrimination in the city, region, or even the whole country where they run their experiment. However, the generalizability of findings to a broader population depends on some requirements being met (for a general discussion of validity issues in experiments: Shadish et al., 2002). First, for valid descriptive results, the level of discrimination has to be the same in the studied and non-studied parts of markets. For this assumption to hold, characteristics that predict the level of discrimination have to be unrelated to the probability of being sampled (Bell and Stuart, 2016). Second, researchers are often also interested in causal relationships with characteristics of the applicants, suppliers or housing markets. For these associations to be externally valid, it would be necessary that one has included all important moderator variables (Shadish et al., 2002).

The gold standard to ensure these assumptions would be to employ a (simple) random sample of tested units (Shadish et al., 2002, ch. 3). However, there exists no official register of housing units that could be used for such purpose. Researchers have instead just relied on random or convenient samples of units that were advertised in newspapers or on internet platforms. In the following, we will first discuss how this might have led to a sampling bias. Second, we review theories on discrimination, with a special focus on the effects of characteristics likely affected by a sampling bias (size of supplier, advertisement time). In sum, this allows us to derive specific predictions on how sampling techniques affect the measurement of discrimination.

Table 4.1.: Sampling Techniques in E-Mail Correspondence Tests Published in 2010 – 2019

Study	Country	Nation-wide[a]	Tested ethnicities	Design[b]	Sampling frame[c]	Field time (in months)	Cases (N Apart.)	Discr. rate (in ppts.)[d]
Acolin et al. (2016)	France	y	African, Turkish, Polish, Portuguese/ Spanish	b	s	2	1,800	13.9
Ahmed et al. (2010)	Sweden	y	Arab	w	a	2	1,032	14.0
Andersson et al. (2012)	Norway	y	Arab	b	a	3	950	12.7
Auspurg et al. (2017)	Germany	n	Turkish	w	a	6	637	9.1
Baldini and Federici (2011)	Italy	n	Arab, Eastern European	b	a	4	3,676	15.0
Bengtsson et al. (2012)	Sweden	y	Arab	b	a	3	1,213	7.3
Björnsson et al. (2018)	Iceland	y	Polish	w	a	5	127	7.9
Bosch et al. (2010)	Spain	n	Moroccan	b	s	3	1,809	12.7
Bosch et al. (2015)	Spain	n	Moroccan	b	s	7	1,186	17.1
Bunel et al. (2019)	France	n	Kanak	w	s	5	342	8.6
Carlsson and Eriksson (2014)	Sweden	y	Arab	b	a	5	5,827	10.7
Ewens et al. (2014)	USA	y	African American	b	s	2	14,237	9.3
Hanson et al. (2011)	USA	n	African American	w	s	4	4,728	6.3
Hanson and Santas (2014)	USA	n	Latino	w	s	1.25	3,072	−0.2
Heylen and Van den Broeck (2016)	Belgium	y	Moroccan, Turkish	b	s	2	1,769	18.6
Hogan and Berry (2011)	USA	y	African American, Asian, Arab, Jewish	w	a	4	1,124	2.9
Mazziotta et al. (2015), Study 1	Germany	n	Turkish	b	s	0.25	336	29.3
Mazziotta et al. (2015), Study 2	Germany	n	Turkish	b	s	0.25	456	14.0
Murchie and Pang (2018)	USA	y	African American, Arab, Latino	b	s	1.5	9,672	3.4
Oblom and Antfolk (2017)	Finland	n	Arab	w	a	5	800	13.7

Notes: [a]y: yes, n: no; [b]b: between, w: within; [c]sampling on the level of suppliers (s) or apartments (a); [d]risk difference between the response probability of the majority and the minority applicants in percentage points (ppts.).

4.3.1. Do Sampling Techniques Lead to a Biased Sample of Housing Units?

As shown in Table 4.1, about half of the recent experiments used apartment-sampling, while the other half used supplier-sampling. Technically, the latter represents sampling *without replacement* on the level of suppliers: Only the first drawn unit of each supplier is kept in the sample. As a consequence, offers by large, commercial suppliers are under-represented, whereas offers by small (often private) landlords are probably over-represented compared to the typical search process of a real apartment seeker.[2]

The second reason why the tested housing units might not be fully representative of the housing units real apartment seekers assess is the restricted field period. The tested housing units are necessarily sampled during a limited period, which typically lasts between one week and several months (see again Table 4.1). The shorter this time interval, the more likely it is that units that are advertised for a relatively long time are over-sampled. This length bias is graphically depicted in Figure 4.1. The x-axis of the *Lexis Diagram* shows the calendar time, while on the y-axis there is the time housing vacancies are advertised online. The vectors represent different offers. Field experiments use a cross-section, also called point sampling of these offers (see the shadowed area in Figure 4.1). Only units that are online during this restricted sampling period have a non-zero probability of being sampled. This means that the probability of being sampled is proportional to the time a unit is advertised online. Put differently, in particular units with a high survival time on the platform (long vacancies) are sampled. These units are those where it takes *per se* more time to find a tenant; or where landlords or agencies are especially picky in choosing their tenants. As can be seen from Figure 4.1, the shorter the sampling period, the more severe is this length bias.

One might argue that real apartment seekers also face this bias. They also look for housing only during a limited search interval and do not follow single apartments over time since they disappear from the market. However, as far as the length of the sampling window does not match the typical search interval of real apartment seekers, the sample would nevertheless be biased.

2 A solution to restore a simple random sample would be the use of *design weights*, i.e. to weight all units by the inverse likelihood of being sampled, which is here the number of offers that were advertised by the same supplier. However, we are not aware of any studies where this technique was used.

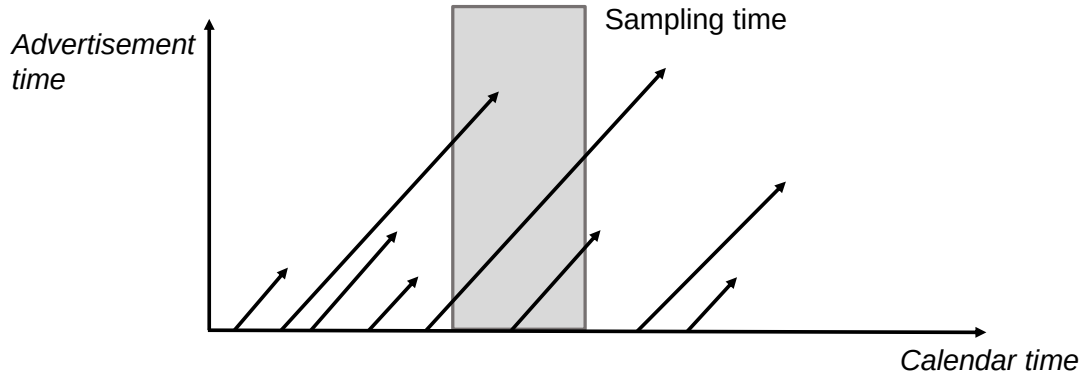

Figure 4.1.: Lexis Diagram: Association Between the Advertisement Time and Likelihood of Being Sampled

Notes: This diagram shows on the x-axis the calendar time, and on the y-axis the time housing offers are advertised online. The different vectors stand for different offers. The height of each vector displays their survival time online (i.e. advertisement time). As can easily be seen, cross-sections mainly catch offers with a long advertisement time (i.e. low success rate). This phenomenon is called *length bias*. Source: Authors based on van Es et al. (2000, 307).

4.3.2. Does Discrimination Vary with the Size of Suppliers or the Length of Advertisements?

There are two main theories on discrimination: *First,* theories on *animus discrimination* or *tastes for discrimination* (e.g. Becker, 1957) assume that suppliers of apartments discriminate because contact with a disliked minority group, such as foreigners, would cause them a psychological disutility. To avoid this disutility, the suppliers are seen to bear economic losses, e.g. in the form of longer vacancies and forgone rental payments, until a tenant of the preferred majority group is found.[3] *Second,* theories on *statistical discrimination* (Aigner and Cain, 1977; Arrow, 1973, 1998; Phelps, 1972) assume that actors discriminate to lower problems of missing information. On the housing market, it is for instance difficult to predict to what extent possible tenants will provide stable rental payments. The ethnicity of apartment seekers could be used as a cheap proxy for this unknown characteristic, as migrants are known to have in general lower earnings and also a higher risk of being in unstable employment. However, in contrast to animus discrimination, these

3 A variant of this animus discrimination is *discrimination due to implicit attitudes*. Following these theories, discriminating actors act mainly unintentionally, based on quick (lexicographic) heuristics (Bertrand et al., 2005; Quillian, 2006). In doing so, they follow unconsciously negative attitudes or (wrong) stereotypes on social groups, which are for instance revealed by psychological test instruments such as the Implicit Association Test (Greenwald et al., 2003).

theories assume that discrimination helps suppliers *increase* their economic profits.

How are these mechanisms related to the size of suppliers and the time apartments are advertised online? Starting with the *size of suppliers*, there are plausible arguments in both directions that larger suppliers (that offer more housing units) are associated with a lower or higher level of discrimination. A prominent argument for a higher level of discrimination is that only the *big players* have enough market power (or assets) to survive in competitive markets when they engage in costly animus discrimination (e.g. Ashenfelter and Hannan, 1986). However, it might also be the other way round: Large agencies probably only grew large because they did not engage in costly animus discrimination (Becker, 1957; Hellerstein et al., 1998). In addition, large suppliers are typically corporate agencies that act solely as brokers, and therefore have only limited contact with possible tenants (e.g. during the showing, the contract processing, and related correspondence). For this reason, tastes for discrimination certainly matter more for private landlords who start a longer-lasting relationship with their tenants. For similar reasons, corporate agencies can also be expected to have less impulse toward statistical discrimination, as it is not them but only apartment owners who would be affected by outstanding rental payments. Also the more formal and rule-driven processes within large agencies are expected to lower the risks of discrimination (Biddle and Hamermesh, 2013).[4] Indeed, there is already some evidence that real estate agents discriminate significantly less against minority applicants than private landlords (Flage, 2018). *We therefore expect that the typical over-sampling of small suppliers leads to an over-estimation of the overall extent of discrimination.*

Regarding the *time apartments are advertised online*, again different mechanisms come into play. First, suppliers without tastes can rent out their apartments more time-efficiently: They are *by definition* less picky, and hence should be able to find a tenant in a shorter time interval than landlords with tastes for discrimination. This means that apartments with long advertisement times are more likely those that are offered by discriminating actors. However, apartments with low success rates could also indicate a tight market. This could be a regional market with many

4 This is also because the larger suppliers are more likely monitored and sanctioned for discrimination. For instance, legislation in Germany explicitly prohibits selecting tenants based on ethnicity only for larger suppliers that rent out apartments beyond their or their close relatives' place of living (see the German General Act on Equal Treatment [Law to implement the European Directive on the realization the principle of equal treatment], § 19 (`http://www.ilo.org/dyn/natlex/natlex4.detail?p_lang=en&p_isn=77201`).

vacancies, and at the same time only few apartment seekers; and/or market segments that have a relatively low demand (such as apartments that are relatively overpriced, offer few amenities and/or are located in poor neighborhoods). A standard prediction made by economic theories is that in tight markets mainly actors who act cost-effectively survive the competition, which are actors with no tastes for discrimination (Becker, 1957). A more realistic assumption is that landlords or agencies do not wait until they are driven out of the market but rather adapt their (discriminating) behavior to the market situation (Fernandez and Campero, 2014). Put differently, discriminating actors might not discriminate at any price, but only as far as the utility gained by discrimination outweighs the opportunity costs in the form of longer vacancies (Biddle and Hamermesh, 2013). If this is true, suppliers will show in general less willingness to discriminate, the more difficult it is to fill their dwelling, and hence the longer they have already advertised their apartment.

Therefore, one might expect two opposing effects related to the length bias. On the one hand, over-sampling units with low success rates could mean that one samples mainly the apartments of suppliers who discriminate. On the other hand, the lower the success rate, the less likely it is that actors still follow their tastes for discrimination. The net effect that results from these two countervailing mechanisms is difficult to predict. We therefore only *expect that the level of discrimination varies with the length of advertisement time*, and we will also explore the *possibility of non-linear patterns*.

4.4. Data

4.4.1. Sample of Housing Units

Our data were collected in a nationwide field experiment in 2015 in Germany, with the tested housing units being advertised on the most prominent online platform for housing units at that time.[5] On this platform, a large number of housing advertisements of both private as well as corporate suppliers were listed. The field experiment took place in two sampling periods, each of five days, in spring and winter 2015 (May 4^{th} – May 8^{th}, and November 30^{th} – December 4^{th}).

On each day in both sampling periods, a random sample of 500 advertised rental apartments was drawn, which resulted in a total sample size of $5,000$ rental apart-

5 More information on the platform and used web scraping methods is available on request.

ments. The sampling procedure was restricted to apartments with $2-4$ rooms. Furthermore, each supplier was tested only once for each sampling period (i.e. we used a sampling without replacement on the level of suppliers). In doing so, we employed a typical sampling strategy for field experiments on the housing market (see again Table 4.1).

4.4.2. Experimental Design

The main treatment of the experiment was the ethnicity of the two male applicants (the gender of applicants was held constant to increase the statistical power regarding the effect of ethnicity). Each supplier got one inquiry from a Turkish and one inquiry from a German applicant (within-design). Both inquiries were sent in random order with a lag of about one hour between the two applications. Ethnicity was signaled by the names, included both in the signatures and the e-mail addresses of the applicants.[6] Turkish immigrants form the largest immigrant group in Germany (German Statistical Office, 2017) and were also tested in preceding correspondence studies (Auspurg et al., 2017; Mazziotta et al., 2015).

Besides their ethnic background, several other applicant characteristics were experimentally varied, such as (the extent of) information on the applicants' educational level (signaled by different occupations) and employment status.[7] The occupation was varied on three levels: There was either (1) no information; or the applicant indicated (2) an occupation that requires vocational training (indicated by six different occupations, e.g. the applicant mentioned that he is currently working as a nurse or electrician); or he mentioned (3) an occupation that requires a university degree (again signaled by six different occupations, such as working as a medical doctor or architect). In case of statistical discrimination, one would expect

6 For both German and Turkish applicants we selected 30 male names that were common and supposed to not signal a specific socioeconomic status, birth cohort, or invoke any other idiosyncratic associations. German names included, for example, Benjamin Buchholz, Maximilian Böhme, and Andreas Engelhardt. Turkish names were, for example, Volkan Sengül, Orhan Simsek, and Erol Tasdemir. All e-mail addresses followed the same format (name.lastname@provider), using a range of common providers, such as aol.de, gmail.com, and gmx.de.

7 We also varied some further information, such as whether there is information on the household income and family status or not. For the analysis presented later we show exemplary results on the education and employment status. The main conclusions do not change when including further experimental treatments, but focusing only on these two main predictors for discrimination helps to keep the presentation and discussion of results more clear cut (results on the other experimental treatments are available on request).

that the gap in the response probabilities declines once information on an occupation requiring a high educational level is provided, as these occupations typically offer an especially high salary and job security (for a more detailed discussion and evidence see e.g. Auspurg et al., 2017, 2019b). The employment status was varied on four levels: (1) no information (to be again able to test for statistical discrimination); (2) permanently employed; (3) self-employed; or (4) working in the public sector. Being self-employed was thought to signal an insecure income. Being permanently employed or working in the public sector, in contrast, was thought to represent a particularly high level of income security due to the particularly high level of job security in the public sector in Germany.

The characteristics of the applicants were completely crossed with each other based on a *D*-efficient experimental design (for details: Auspurg and Hinz, 2015) that minimizes the correlations between the different treatment variables and at the same time maximizes their variance. Such a design allows for a maximum level of statistical power to estimate the impact of all treatments. Note that the ethnicity was always varied within the tested apartments, while other treatment variables could be the same or different for the two applications sent to one apartment. In order to minimize the risk of the experiment being detected, different wordings for the two inquiries were used, such as different salutations or orderings of the information displayed in the e-mails. The resulting pairs of e-mails (one always being of a Turkish and one of a German applicant) were randomly allocated to the sampled apartments. We carefully checked whether the randomization worked (i.e. approved that applicant characteristics were not correlated with any characteristics of the suppliers or regions where the experiments were done). We also assured that the different text versions did not evoke any idiosyncratic response patterns.[8] Figure 4.2 shows an example inquiry, with the experimentally varied applicant characteristics being underlined.

4.4.3. Market Data

We collected official statistics on the regions where the apartments were located. Besides that, we combined our experiment with market data captured on the internet platform itself. While the experiment took place in two sampling periods (in May

8 The maximum correlation of a text version with the observed response pattern (i.e. discrimination) was $r = 0.025$, $p = 0.17$.

Dear Ms./Mr.,

I am highly interested in the advertised apartment. My name is Volkan Sengül and I am permanently employed as an electrician. I am looking for an apartment for me and my family. I would be very grateful if you could offer me a showing and information on similar offers in the neighborhood.

Kind regards,

Volkan Sengül

Figure 4.2.: Sample Inquiry (Translated Version, Experimentally Varied Dimensions Are Underlined)

Notes: In other e-mail variants, additionally information on the applicant having a partner or family and his income was provided. The whole list of variations and text phrases is available on request.

2015 and November/December 2015), each of five days, the platform was observed for a whole year, covering at least one month before and after the two sampling periods. Information on all advertised apartments was collected daily between March 2015 and February 2016, exporting each active advertisement with an automated web-scraping routine.[9] The outcome is a database that identifies the spells in which the apartments were advertised on the platform. All in all, we gathered spell data on $1,087,541$ advertised apartments.

These data allow us to calculate the information we are interested in: the size of suppliers, measured by the number of offers they advertised during our one-year observation period; and the time offers were (already) advertised online (until we sent our e-mails). Suppliers that did not report any company name were coded as private landlords. For private landlords, we could not observe the exact number of offers belonging to the same supplier, because no distinct information was available on the platform.[10] Therefore, all private landlords were coded to have only one advertisement by design. Due to the high advertisement prices and the differential pricing offered by the online platform, this seems plausible: All suppliers could set up a corporate account that offered discount rates when having several offers, so

9 Due to technical problems there were some gaps of two to nine days where we had no observations, spreading over the whole observation period (except for the experimental periods). This could have led to some censoring of advertisement times.

10 Some private suppliers only indicated their first or second name, which does not allow for a valid identification of single suppliers.

landlords with multiple offers likely chose this option.

The advertisement time was measured in days. Short interruptions in advertisement times of up to 14 days were 'smoothed' (i.e. seen as one joint advertisement interval), as it is very unlikely that an apartment was rented and offered again within a time window of only two weeks. It is more plausible that it is still the same vacancy: Taking an apartment offline, organizing some showings, and putting it online again if there was no adequate tenant represents a cost-effective renting-strategy for landlords. Besides using a very long observation window, in some cases a left- or right-censoring of the advertisement time occurred (i.e. advertisements were already or still online when we started or finished monitoring the internet platform). We will provide results based on observations without censoring; robustness checks that also included censored advertisement times did not change any substantive conclusions (results are available on request).

We present results on apartments located in Western Germany. We decided to exclude Eastern Germany because the rental market as well as attitudes toward foreigners still strongly differ across both parts of Germany. To adequately capture these heterogeneities, one would have needed quite complex regression techniques (Auspurg et al., 2019a). However, it is noteworthy that our substantive results still hold when including Eastern Germany. After deleting cases with missing observations on, for example, the advertisement time (mostly caused by left-censoring), $2,992$ tested apartments (of formerly $N = 3,932$) remained in the sample. This number corresponds to $5,984$ ($= 2,992 \times 2$) e-mail applications, half sent by a Turkish and half sent by a German applicant.

4.4.4. Identification Strategy

To see if there is the expected sampling bias in the size of suppliers and length of advertisements, we first contrast descriptive statistics on the sampled housing units (the *sample*) with statistics on the full population of housing units within the one-year observation window (the *market data*). To see if a possible sampling bias in these variables affects the measurement of discrimination or conclusions on theories on discrimination, we will then explore whether a) the measured level of discrimination and b) its association with applicants' characteristics (such as them providing more or less information) depends on the two apartment characteristics (size of supplier and advertisement time).

To *measure discrimination*, we follow standard procedures in the literature and look at the quantity of responses. Getting no response is definitely a rejection and in nearly all cases the response to an e-mail inquiry is a positive one (i.e. an invitation to a showing).[11] In our analyses, we will look at the apartment level and contrast the following three outcomes j:

(1) *Equal treatment ($j = 0$)*: Both the Turkish and German applicant get a response or both get no response.

(2) *Discrimination against the Turk ($j = 1$)*: Only the German (but not the Turk) gets a response.

(3) *Discrimination against the German ($j = 2$)*: Only the Turk (but not the German) gets a response.

The percentage of cases falling in category 2 (3) is the so-called gross discrimination rate against the Turkish (German) applicants. In addition to this, authors often report the net discrimination rate of minorities, which is defined as the difference between both discrimination rates (i.e. 'discrimination against the Turk' minus 'discrimination against the German;' for a detailed discussion of these discrimination measures see Wienk et al., 1979, p. 18). In our analyses, we will only explore the gross discrimination rates (i.e. the absolute outcomes 2 and 3). Looking at the gross discrimination rates allows for more fine-grained insights than when looking at the net discrimination rate that summarizes both outcomes (note that our results are, however, robust to the alternative net-measurement of discrimination). In addition, this analysis strategy mirrors typical analyses used in the literature (see e.g. Ross and Turner, 2005).

To see how both gross discrimination rates vary with the characteristics of interest, we first use non-parametric estimations (LOESS smoothers: Cleveland, 1979) that help to explore possible non-linear associations. In addition, we use multiple multinomial logistic regressions to contrast the three different outcomes j: Equal treatment is used as the reference category ($j = 0$; for details on multinomial regression models see Greene, 2012, p. 763). Against this reference category of equal

11 Results including qualitative information on the responses likely suffer from lower reliability, as the quality of responses is difficult to code (the subtle forms of discrimination are more difficult to uncover; see Hanson et al., 2011). In our case, including qualitative information led to quite similar substantive conclusions (results on these robustness checks are available upon request).

treatment, we estimate the likelihood of gross discrimination against the Turkish applicant ($j = 1$) and gross discrimination against the German applicant ($j = 2$).

Equation (4.1) shows this regression model. *Logit* specifies the log-transformed odds of the two discrimination outcomes $j = 1$ or $j = 2$ against the reference category of equal treatment $j = 0$. i is an index for the different tested housing units ($i = 1, \ldots, N_{\text{apartments}}$). In all models, we include some control variables C that are known to influence the level of discrimination (percentage of foreigners in the county; apartment located in a city yes or no). T are the treatment variables in the form of the characteristics of the applicant besides their ethnicity, while A are the apartment characteristics of main interest, i.e. the size of the supplier as well as the time the apartment was advertised until treatment (i.e. until the e-mails were sent). Due to the skewed distribution of these variables, they enter the regressions in logarithmic specifications. Positive (negative) regression coefficients mean that the odds of the outcome j is increased (decreased) compared to the reference category. For our research aim the coefficient β_{Aj} in equation (4.1) is of most interest: A significant coefficient β_{Aj} would suggest that sampling bias in the tested housing characteristic A translates to a biased estimate in the level of discrimination (outcome j).[12] Positive (negative) effects mean that larger suppliers/longer advertisement times go along with higher (lower) odds of discrimination.

$$\begin{aligned} \text{Logit}(Y_i = j) = \beta_{0j} + \beta_{Aj} \ln A_i + \beta_{Tj} T_i + \beta_{Cj} C_i, \\ j = 0, 1, 2; \quad i = 1, \ldots, N_{\text{apartm.}} \end{aligned} \tag{4.1}$$

In a second step (equation 4.2), we additionally include interaction terms between the treatment (applicants' characteristics) and apartment variables ($T \cdot \ln A$) in order to see if the sampling bias also affects the estimated effects of the treatment variables (and related tests of discrimination theories). A significant coefficient β_{TAj} of the interaction term would suggest that sampling bias translates to a biased assessment of the effects of treatment variables: Depending on the sample composition in regard to A (i.e. size of suppliers, advertisement times), different treatment effects would

12 Multinomial regressions estimate different regression coefficients for the different outcomes. In our case this allows to see whether the explaining factors for the discrimination against the Turk versus German (and possible measurement bias herein) differ.

be estimated.

$$\begin{aligned} \text{Logit}(Y_i = j) = \beta_{0j} + \beta_{Aj} \ln A_i + \beta_{Tj} T_i + \beta_{TAj} T_i \cdot \ln A_i + \beta_{Cj} C_i, \\ j = 0, 1, 2; \quad i = 1, \ldots, N_{\text{apartm.}} \end{aligned} \tag{4.2}$$

The full regression models are shown in the Appendix (`http://dx.doi.org/10.6084/m9.figshare.9890801`). In the main text we will provide only visual representations of the effects of main interest. Multinomial regression coefficients give the effect on the logit and are difficult to interpret. We therefore present average marginal effects (AMEs), which indicate the mean increase in the probability of an outcome j (percentage points when multiplied by 100) that is caused by a marginal increase of the variable of interest, when averaged over all observations. All analyses are done with the statistical software Stata version 15 (StataCorp., 2015). For the local polynomial plots, the Stata procedure *lpolyci* was used, and the coefficient plots were created with the user-written Stata ado *coefplot* (Jann, 2014).

4.5. Results

4.5.1. Is there a Sampling Bias?

Table 4.2 shows descriptive statistics on the sample used in the experiment and on all offers advertised during our one-year observation window (see Section 4.4.3 for details). As expected, there were substantial differences. First, regarding the size and kind of suppliers the expected over-sampling of small, private suppliers occurred: In the sample, about half of all suppliers were private landlords (53%), while only 26% of all suppliers active on the internet platform were private. In particular the number of offers per supplier differed drastically between the sample (mean 22; median 1) and market data (mean 1,246; median 34). This huge difference was especially caused by the supplier-sampling: Even the largest supplier, although advertising 15,810 apartments in our one-year observation window, *by design* was sampled only once.

Second, there was also strong evidence for the expected length bias. In the sample, the mean time an offer was advertised[13] on the platform was 52 days, which is more than twice the time found for the complete market data (22 days). In contrast,

13 This is the overall time an advertisement was online. In the models presented later on we will use the advertisement time until experimental treatment (i.e. until we sent our e-mails).

Table 4.2.: Descriptive Statistics on the Sample and Market Data

	(1) Sample Drawn for the Experiment		(2) Market Data: All Offers in the One-Year Observation Period	
	Mean	SD	Mean	SD
Supplier characteristics				
Private suppliers[a]	0.53		0.26	
Number of offers per supplier	22.35	203.10	1,246.13	3,303.80
Apartment characteristics				
Advertisement time (days)[b]	51.99	44.81	22.19	28.63
Size (sqm)	83.50	26.52	82.75	24.46
Prize (per sqm)	8.19	3.12	8.11	2.79
N	2,992		695,458	

Notes: [a]Coded as 'private' when there was no information on a company name (see Section 4.4.3). [b]Due to censored information on the advertisement time, we can calculate this statistic only based on 2,598 cases in our experimental data and 575,950 cases in the market data. In the analyses presented later on, we alternatively use the information on the advertisement time until our experimental treatment took place (i.e. we sent our e-mails), which allows for a slightly higher number of observations (N = 2,992 apartments).

other characteristics of the apartments, like the cost of the rent or the size, differed only slightly between both data sources.

Because of the large skewness of the advertisement time as well as the number of advertisements offered by a supplier, both variables were log-transformed. In this logarithmic specification, large differences between sampled units and the whole population (market data) persist (see the kernel density estimates provided in Figure 4.3). Furthermore, it can be seen that especially the largest suppliers in the right tail of the market data were under-represented in the experimental data in favor of very small suppliers (with mainly only one offer; see the kernel density estimate in the left panel). For the advertisement time, the distribution in the experimental data is shifted to the right, showing an over-representation of advertisement times of at least 20 days (re-transforming the logarithmic value 3 gives $e^3 = 20$).

To see if the observed sampling bias does not only hold for our unique sample of housing units, we also run simulations of different samples (see Table 4.A1 in the Appendix). Simulating both sampling frames (supplier versus apartment level) in combination with different sampling periods (ranging from one week to six months),

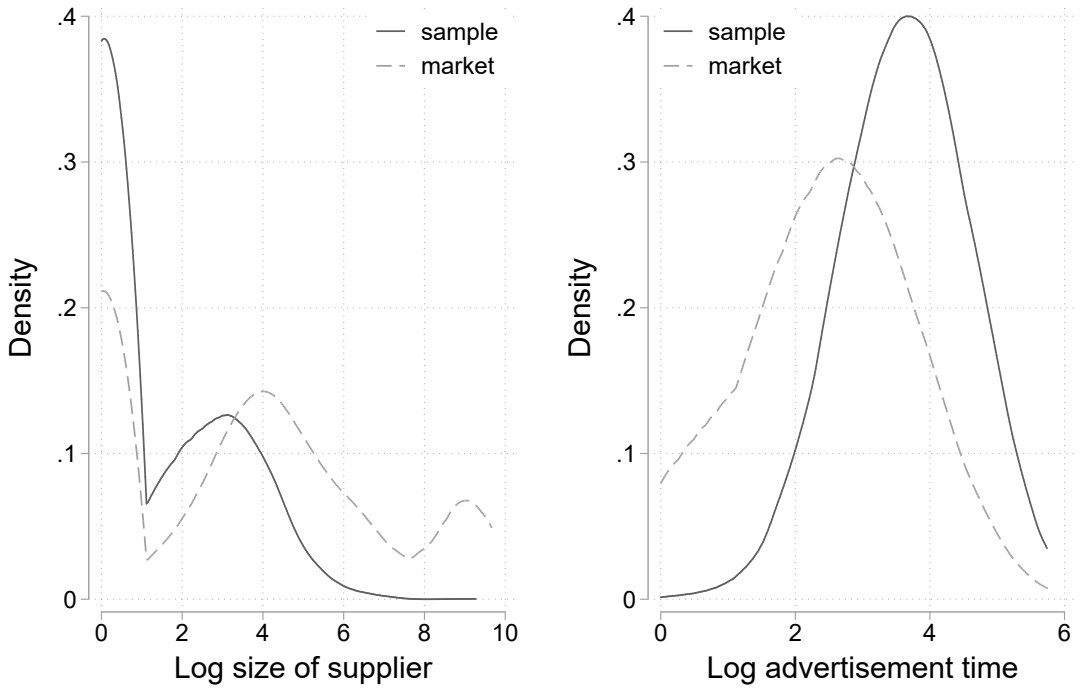

Figure 4.3.: Kernel Density Estimate of the Size of Supplier (Left Panel) and Advertisement Time (Right Panel) by Data Source (Sample vs. Market Data)

Notes: Size of supplier: number of offers per supplier. Advertisement time: overall time (in days) the apartment was advertised on the internet platform. Both graphs were produced using the Stata command kdensity. For the number of cases see Table 4.2.

we always found the expected biases: Sampling on the supplier level led to an underrepresentation of large suppliers, while with longer sampling periods the observed advertisement time decreased. This makes us confident that *there is indeed a strong sampling bias in terms of over-representing small suppliers and offers with long advertisement times.*

4.5.2. Does Sampling Bias Affect the Level of Discrimination?

Table 4.3 shows a cross-tabulation of responses to the Turks' and Germans' applications. One can see that in most of the tested apartments ($81.8\% = 32.6\% + 49.2\%$) the suppliers treated the Turkish and the German applicant equally: In 32.6% of the cases they answered neither of the two e-mails; and in 49.2% of the cases they

responded to both e-mails. These cases will form the reference category in our multinomial regression analyses (see Section 4.4.4). In the remaining cases, unequal treatment or *discrimination* occurred. The observed *gross discrimination rate against the Turks* is 14.4% (in 14.4% of all cases only the German applicant got a response). The *gross discrimination rate against the German* is 3.8% and hence quite lower (only in 3.8% of all cases the Turkish applicant alone got a response). Taken together, this means that there was a net discrimination against Turks of $14.4 - 3.8 = 10.6$ percentage points. This discrimination rate is close to that reported in other field experiments on Germany, and it also comes close to the risk difference in majority and minority response probabilities reported in our literature review in Section 4.2.[14]

Table 4.3.: Response Patterns and Resulting Discrimination Rates

		German applicant (G)		Total
		No response	Response	
Turkish applicant (T)	No response	975 (32.6%)	430 (14.4%)	1,405 (47.0%)
	Response	114 (3.8%)	1,473 (49.2%)	1,587 (53.0%)
Total		1,089 (36.4%)	1,903 (63.6%)	2,992 (100.0%)

Notes: This table shows the obtained response patterns for the 2,992 apartments tested in Western Germany with full information on the supplier size and advertisement time until treatment (i.e. until we sent our e-mails). The 14.4% (3.8%) of cases with only a response to the German (Turkish) applicant define the gross discrimination rate against Turks (Germans).

Do the discrimination rates vary by the size of supplier and/or length of advertisement times? Before switching to regression analyses, we present in Figure 4.4 non-parametrical estimates that rely on fewer assumptions and help to identify possible non-linear patterns. Due to the very skewed distribution of both independent variables (see Table 4.2) we use logarithmic specifications, and restrict the analyses to values with sufficient observations for stable estimates. The shadowed areas display the 95% confidence intervals. In the left panel one can see that with an increasing size of suppliers the gross discrimination against Turks declines: From

14 These risk differences are recommended to be used in meta-analyses, as reported in Section 4.2. For the research aims in this paper, however, it seems more reasonable to rely on the simpler (and easier to interpret) gross discrimination rates.

16.0% for suppliers that offer only one apartment (which is about half of our sample) to 12.5% for the largest suppliers (that offer up to ln (5), which is about 150 offers). Given a mean discrimination rate of 14.4%, this decline by more than four percentage points might be seen as substantial in effect size; it is, for example, of a similar level as differences found across countries (see Table 4.1 and Auspurg et al., 2019b). Unequal treatment against the German applicant in contrast remained constant at 4% over the whole range of supplier sizes. Therefore, the net discrimination rate, indicated by the distance between both discrimination rates, also declined with the increasing size of suppliers.

For the advertisement time (time online until experimental treatment, i.e. until we sent our e-mails), a similar pattern was observed (see the right panel). Unequal treatment against the Turkish applicant decreased from around 17% for very short advertisement times (1 day) to about 13% for very long advertisement times (of about 150 days). In contrast, unequal treatment against the German applicant was always around 4%. Hence, again, there were marked differences in the observed level of discrimination. However, due to the very large confidence intervals, the variations in the level of discrimination probably do not reach statistical significance.

To back up these results, we estimated parametric multinomial logistic regressions. The full models with all logit coefficients and information on the model fit can be found in the Appendix (Table 4.A.2, which shows regressions with a stepwise inclusion of covariates). Figure 4.5 only displays the effects of the apartment characteristics of main interest, the size of the supplier and advertisement time, while controlling for applicant characteristics besided ethnicity and some further control variables (see the Figure notes). Effects are reported as AMEs. Looking first at the 'discrimination against the Turkish applicant' (left panel), one can see that (compared to the reference of 'equal treatment') the likelihood of this outcome lessens with the size of suppliers (by on average 0.92 percentage points if the supplier size is increased by one log-unit: AME $= -0.0092$; $p < 0.01$; see Table 4.A.2 and 4.A3 in the Appendix for the exact estimates).[15] This mirrors the descriptive results reported so far.[16] When switching from the smallest (only one offer) to the largest

15 Table 4.A.2 shows the logit estimates that were chosen for reporting the models with interaction effects, as interaction effect estimates are misleading when linearizing nonlinear models by AMEs (Norton et al., 2004). Table 4.A3 shows in addition the AMEs presented in the Figures.

16 When controlling for 'private landlord' the effect was no longer statistically significant, meaning that the effect was mainly driven by the higher discrimination related to small private landlords. In the analyses shown here, we do not control for private versus commercial land-

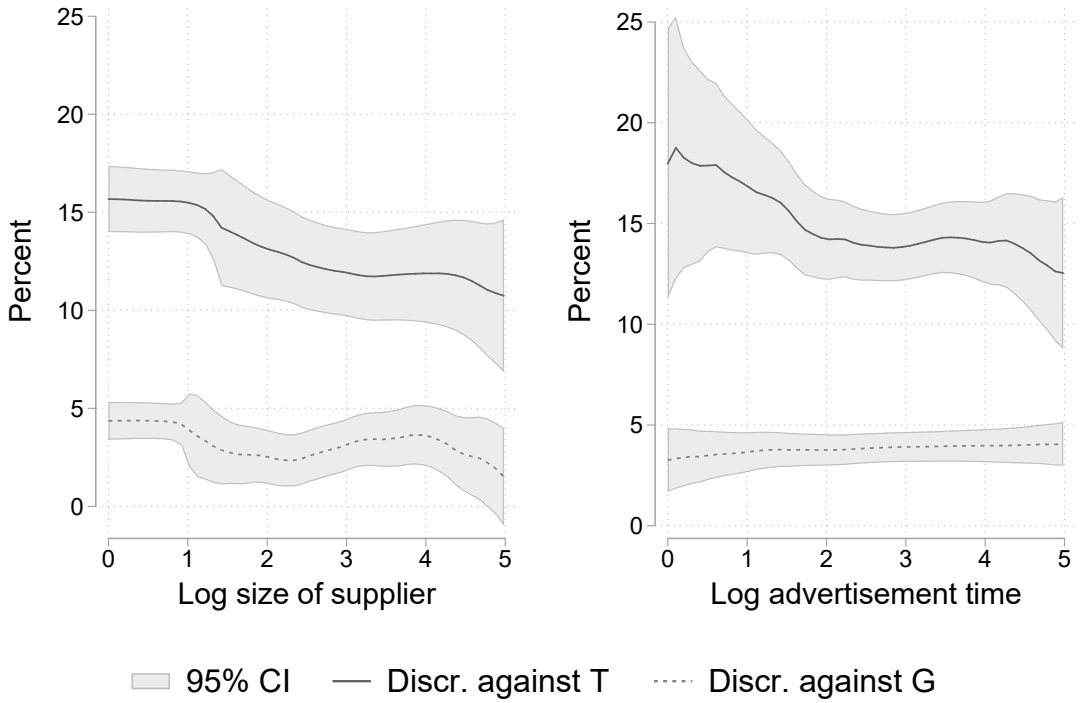

Figure 4.4.: Local Polynomial Smooth Curves of the Discrimination Rates Against the Turkish (T) and German (G) Applicant Over the Size of Supplier (Left Panel) and Advertisement Time (Right Panel)

Notes: The gray areas show the 95% confidence intervals. G: German; T: Turkish. The left panel shows the log-number of offers per supplier on the x-axis, the right panel the log-number of days an apartment was already advertised online until treatment. Both figures are based on $N = 2,992$ apartments. The figures were produced using the Stata command *lpolyci*.

suppliers (∼150 offers), the discrimination rate is predicted to decline on average by 4.61 percentage points (as $-0.0092 \cdot \ln(150) \cdot 100 = -4.61$).

The time an offer was advertised online (until treatment) also showed a negative effect on discrimination against the Turkish applicant, but this effect was not statistically significant (the confidence intervals overlap with the zero-line). For discrimination against German applicants, no substantial effects are found. All these results were robust to other linear and non-linear specifications (the reported log-specification of both variables provided the best model fit). *All in all, we can conclude that over-sampling small suppliers (which is mainly caused by a supplier-sampling) leads to a small, significant over-estimation of discrimination against mi-*

lords to prevent an 'overcontrol bias' caused by the inclusion of mediator variables (Elwert and Winship, 2014).

nority applicants; while short field periods (i.e. over-sampling long advertisements) tends to lessen the observed discrimination rates, but only to a non-significant degree.

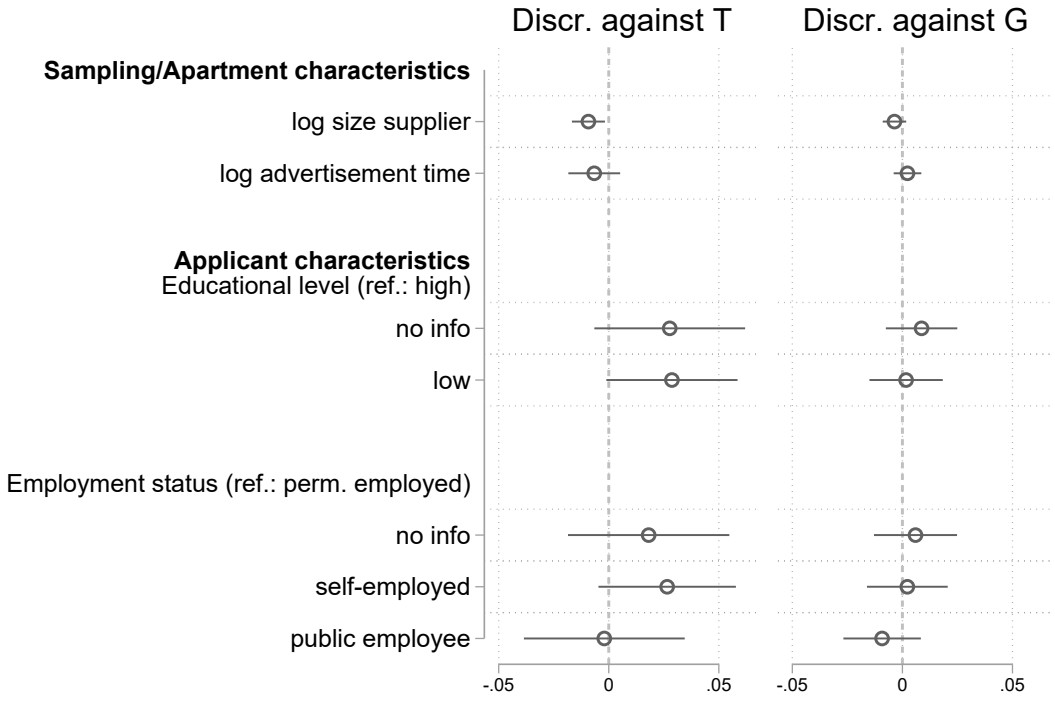

Figure 4.5.: Multinomial Logistic Regressions, AMEs of Predictors with 95% Confidence Intervals for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment')

Notes: Displayed are AMEs. For the outcome 'Discrimination against the Turkish applicant' ('Discrimination against the German applicant') effects of the characteristics of the Turkish (German) applicant are shown. Characteristics of the other applicant are controlled, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table 4.A.2 in the Appendix for logit estimates for all variables and Table 4.A3 in the Appendix for the AMEs displayed in this Figure. Results are based on $N = 2,992$ apartments.

Figure 4.5 also reports the effects of some applicants' characteristics. For the Turkish applicants (see again the left panel), one can see that both providing no information or indicating a low educational level (signaled by the occupations mentioned in the e-mails) tended to increase the probability of being discriminated against. The latter effect is significant at the 10% level ($p = 0.065$). Regarding employment status, no information also tended to increase the risk of discrimination against Turks, and there was also a tendency that self-employed Turks were more

likely discriminated against compared to the reference of permanently employed applicants (this effect was again significant at the 10% level: $p = 0.075$). In regard to discrimination against the German applicant, again no substantial effect was found. In sum, this means that there was only weak evidence for statistical discrimination against Turks: Information that signals a higher or more stable income tended to lower the risk of discrimination; but overall the effects were not strong enough to reach statistical significance.

4.5.3. Does Sampling Bias Affect the Effects of Other Treatment Variables?

In a final step, we are interested in whether sampling bias affects the measurement of the effects of applicants' characteristics (treatment variables besides ethnicity). To find out, we have to see whether the effects of these variables are moderated by the size of suppliers or advertisement times.

To ease interpretation, and also to be able to observe possible non-linear patterns, we present our results as split sample analyses by three strata of the size of supplier and length of advertisements. For the size of suppliers, we contrast the 56% of suppliers with only one offer with two other strata of about equal size (each containing about 22% of suppliers: 2-17 offers; and 18-10,684 offers); for the length of advertisements, we contrast the three terciles (1-12 days; 13-29 days; 30-245 days). In the Appendix, we provide in addition pooled regression analyses that include the interaction terms of treatment and apartment variables ($T \cdot \ln A$), which allow testing of whether moderation effects by apartment characteristics are statistically significant (Table 4.A.2). In the following, we only show results on the discrimination against Turks, where we found so far the strongest effects for treatment variables; results on discrimination against Germans are provided in the Appendix (Figure 4.A1 and Table 4.A5).

Figure 4.6 shows the results. For the size of the supplier (top panel), no clear pattern emerges: The effects of educational levels seem to be somewhat stronger for larger suppliers, but for employment status it is the opposite (i.e. smaller effects for larger suppliers). All these differences across samples are probably only due to random variations. This is supported by the non-significant interaction of both treatment variables with the size of suppliers in pooled regressions (see Table 4.A.2 in the Appendix).

Regarding the split samples by advertisement time (bottom panel), it is noteworthy that there are for nearly all treatment variables the strongest effects in the stratum with very short advertisement times (up to 12 days). In this sample based on relatively short (and hence new) vacancies, the effects of several treatment variables reach statistical significance. One can, for example, observe that for a low (compared to high) educational level the risk of being discriminated against is significantly higher (by 7.7 percentage points, AME = 0.077). With the other two samples one would, however, conclude that educational background does not make a significant difference. Similar patterns emerge for employment status. These observations suggest that very short field periods (that mainly catch the long advertisement times) somewhat underestimate the incidence of statistical discrimination (i.e. discrimination depending on the amount and kind of information on the applicants). The observation is also in line with our assumption that landlords become less picky when it takes more time to fill their apartments. However, pooled estimates with interaction effects indicate that the differences across strata are not statistically significant (see again Table 4.A.2 in the Appendix). In a practical sense, the results nevertheless suggest that different samples can lead to quite different conclusions—although some of the differences are probably simply caused by sampling error (we return to this in our discussion). Note that for the discrimination against Germans, no remarkable patterns or statistically significant differences were found (see Figure 4.A1 and Table 4.A5 in the Appendix). *Overall, we conclude that sampling techniques and in particular the length of field periods have some impact on the estimated treatment effects; however, in our case study the moderation of treatment effects by size of supplier or advertisement time is too small to reach a statistically significant level.*

4.5.4. Robustness Checks

We presented analyses on the apartment level, where we estimated multiple multinomial regressions to predict both the likelihood of discrimination against Turks and Germans against the reference category of equal treatment (i.e. both getting a response or both getting no response). Substantive results are robust to using an alternative reference category (both getting no response; while treating 'both getting a response' as a separate outcome).

An alternative approach used in discrimination research is to use the e-mail in-

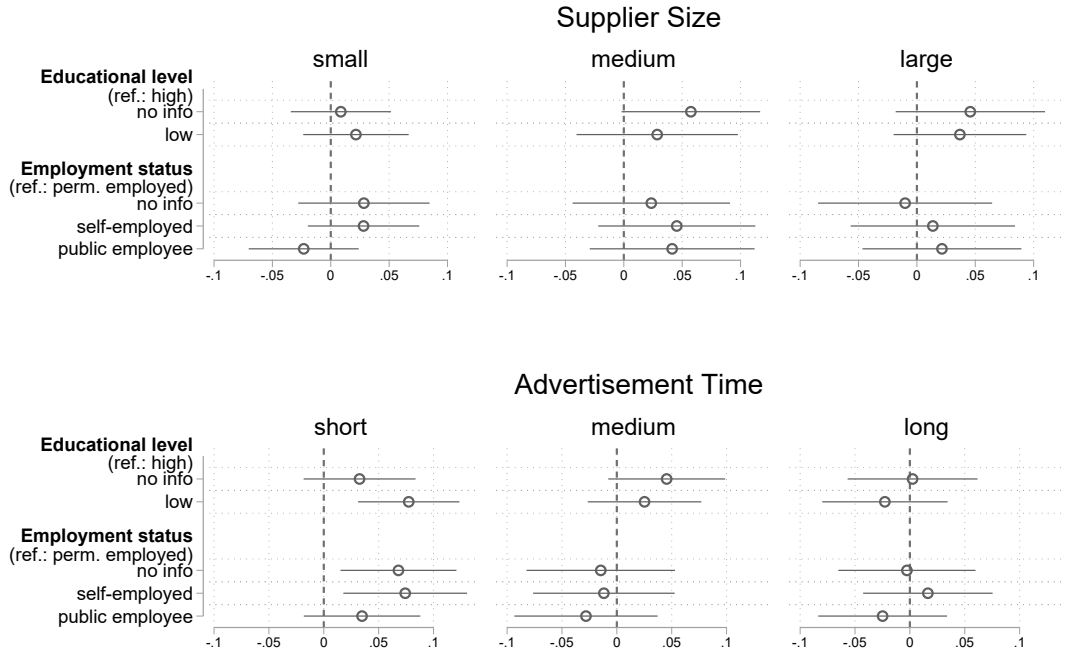

Figure 4.6.: Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time

Notes: Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); (top panel); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment (bottom panel). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table 4.A4 in the Appendix for the exact estimates. All estimates are based on at least $N = 658$ observations.

quiries as units of analysis and to estimate regressions with a dichotomous dependent variable (is there a response, yes or no). In this approach, the occurrence of discrimination is observed by a significant main effect of Turkish ethnicity. As pairs of e-mails are nested within the same apartment, multilevel regressions are appropriate (random effects regressions). When using this alternative approach (where we tested effects on the level of discrimination by bivariate interactions of Turkish ethnicity with supplier size/advertisement time, and the moderator hypotheses by three-way interactions of the Turkish ethnicity, applicants' characteristics, and supplier size/advertisement time), the main conclusions were very similar: There

were again variations in the level of discrimination with a maximum of four percentage points difference in the levels (again the level of discrimination was found to be higher for small suppliers and short advertisement times), and treatment effects differed marginally by these moderator variables. Besides similar effect sizes to the ones reported in our results section, none of these variations reached statistical significance at the 5% level. This is probably due to the much lower statistical power when testing interaction effects instead of main effects (Cohen, 1988, pp. 367–369). Nevertheless, these robustness checks found the same direction and effect sizes.

To ensure that we did not overlook non-linear effects, we not only employed non-parametric analyses (LOESS curves, as presented in Figure 4.4) but also tested different parametric specifications, such as regressions with polynomial terms and categorical variables for the sampling/apartment characteristics.

Finally, we also tested whether weighting regression analyses by the size of supplier in order to correct for the different sampling probabilities leads to different conclusions. Again, this was not the case: Weighting decreased the impact of our market variables on discrimination to a small extent but did not change substantive conclusions on the effects of treatment variables.

4.6. Summary

The aim of our study was to analyze how sampling strategies affect the external validity of field experiments. Combining a large-scale field experiment on ethnic discrimination in the German rental housing market with rich data gathered from the platform where the apartments were advertised, we were able to test how both sampling on the level of suppliers (instead of apartments) and point sampling (i.e. sampling a cross-section of housing units advertised during a limited field period) affects the observed level and mechanisms underlying ethnic discrimination. Although these sampling techniques are frequently used in field experiments, we are not aware of any other study that systematically discussed or analyzed these issues. We summarize our main findings in three points.

First, when comparing our sample to the larger population of housing units where our sample was drawn from, there was clear evidence of a sampling bias: As expected, sampling on the level of suppliers led to a huge over-representation of small suppliers with only few offers. At the same time, there was a strong over-representation of apartments with relatively long advertisement times. These

aspects suggest that standard field experiments over-sample apartments that are offered by private landlords or other agencies with a small market power, and/or offers that stay a relatively long time on the market.

Second, we analyzed how both issues affect the observed level of ethnic discrimination. Empirically, the over-sampling of small/private suppliers particularly tended to increase the amount of discrimination observed in our study. These differences in effect sizes were, however, moderate: Even when contrasting the most extreme strata in our sample (containing only the apartments that were advertised by the smallest or largest suppliers), the found discrimination rate against Turks was within a relatively narrow range of 12.5% to 16.0%. Similarly, long advertisement times only slightly tended to go along with lower discrimination rates (differences in effect sizes were not statistically significant).

Third, we also analyzed whether the sampling bias might have affected the observed risk factors for discrimination. We tested characteristics of applicants (varying information on their educational and employment status) that were commonly used to explore statistical discrimination. We found some moderation effects, especially for the length of advertisements: Effect sizes tended to be a little bit stronger for offers that were advertised only for a relatively short time when our experiment took place, and only for these (new) offers we oberserved effects that point to statistical discrimination. However, none of the moderation effects were found to be statistically significant. All in all, the observed levels and incidence of (statistical) discrimination seemed to be mostly immune to deviations from 'representative', random samples of housing units.

4.7. Conclusions

Putting all these findings together, we conclude that the results of field experiments seem to be remarkably robust against sampling bias in terms of supplier size and advertisement time. Not surprisingly, descriptive findings were found to be more strongly affected than multivariate results: According to our estimates, the level of discrimination might be up to four percentage points larger or smaller, when one studies extreme samples that over-represent specific suppliers or advertisements. Given a baseline discrimination rate (observed on average in our sample) of about 14%, this variation might be classified as substantial. When comparing studies over time or done in different countries, sampling bias might be large enough to obscure

time trends or cross-country differences that were so far found to be of similar size.

We nevertheless conclude that in particular the studied treatment effects in terms of applicants' characteristics (e.g. information on their employment status) were remarkably robust. Given that these variables are of main interest to advance our knowledge on underlying mechanisms (animus or statistical discrimination) and also to advance interventions (Neumark, 2012), this is good news. Given the very similar general patterns found across all strata, and given the large number of variables tested in our study, most differences likely occurred just by chance (i.e. resulted from random sampling error). This rather points to the necessity of consolidating findings with replications based on other (similar or different) subsamples than to strong biases caused by sampling techniques.

Nevertheless, to increase the comparability of findings across experiments, one should try to be mostly transparent on the sampling techniques used. Only this would allow to include this information in future meta-analyses (the existing ones controlled only for few other design variables: Auspurg et al., 2019b; Flage, 2018). A detailed documentation of the sampling procedure would also allow the use of design weights (for the application in surveys see Lavallée and Beaumont, 2015). Re-weighting the sample or using samples on the level of apartments from the beginning on certainly more closely matches the search strategies used by real apartment seekers.

In this context, it has to be stressed that the generalizability of research findings to 'real world processes' does not necessarily require a random sample (Salganik, 2017). Often it is difficult or even impossible to define a clear population and hence sampling frame; or doing so could result in excessive costs (Brewer and Crano, 2014). The results of experiments only do not generalize to other settings when there are differences that moderate the association between treatment and outcome variables. To ensure a high external validity of findings, it is in general important to study not only a broad range of different units, but also treatment variables, outcomes and experimental settings (Shadish et al., 2002). Combining field experiments with market data is herein not only beneficial in terms of safeguarding the external validity of findings, but also in identifying important moderator variables: Observing variance across (market) contexts can provide important insights for advancing theories (Brewer and Crano, 2014). Market data might also simply be sampled from the platforms (without using larger observation windows); for example, one might use

the indicated names of companies to collect further information on the suppliers.[17] Nevertheless, if the goal of research is to describe the 'true' level of discrimination individuals face, one might try to match their search processes as closely as possible to ensure that one really has mirrored the sample composition of all possible moderator variables.

In the last section, we want to discuss how our study points to directions for future research. *First,* and probably most important, we only provided evidence on one single case study in Germany. We tried to allow for more general conclusions by means of simulations to ensure that our observed sampling bias is not only bound to the one, idiosyncratic sample drawn for our field experiment. However, the results nevertheless might not generalize to other countries. In a meta-analysis on the housing market, Germany stands out as having a slightly higher amount of discrimination than is observed in other countries (Auspurg et al., 2019b). This suggests that there are cross-country differences that moderate discrimination, and these differences might also moderate the associations analyzed in our case study. We therefore encourage researchers to repeat similar studies in other countries.

Second, our study was restricted to one online housing market. This might pose an even severe threat to the external validity of findings, as real apartment seekers likely use additional sources of information on vacancies, such as newspapers or social networks. That said, we are not aware of any study that compared samples based on different (offline and online) media. The bias that is caused by focusing only on one specific search strategy or (online) platform is probably more problematic than the bias we analyzed in this study. For instance, it seems plausible that in newspapers and in particular in social networks more units made available by small private (or older) landlords are advertised, and that in particular in social networks tastes for discrimination prevail (for some qualitative evidence on the often very subtle processes of discrimination that take place in personal interactions, see e.g. Krysan and Crowder, 2017). Without systematic research on the effects of sampling frames, one can only speculate on the direction and size of a possible bias. Another fruitful direction for future research would therefore be to study how discrimination

17 When doing so, one has to consider ethical concerns. In addition, one should carefully check the validity of those 'big data'. In particular information provided by the platform itself should be regarded with caution. For example, especially commercial suppliers might pay for having their offers ranked as 'new' even though they have advertised them already for quite some time. Also, 'time drifts' in the way information is provided or idiosyncrasies due to the used algorithms ('algorithmic confounding') represent additional sources of errors for these data (Salganik, 2017).

(and underlying mechanisms) vary with the used (social) media to identify available housing units.

Third, although we collected 'big data' on the internet platform with an observation method that was 'always on' during our one-year observation period (Salganik, 2017), and hence gathered very fine-grained spell data with a longitudinal dimension, the field experimental data represent only a cross-section: Each vacancy was tested only at one specific point in time. This hampers the identification of different mechanisms that might be bound to the length of advertisements. The observed reduction of discrimination over the course an apartment was unsuccessfully advertised online could indicate that discriminating actors gave in to market pressure, and started to make compromises on the preferred attributes of tenants. However, to truly test this mechanism one would need longitudinal data also on the experimental side, i.e. testing single apartments several times over the course they are advertised online.[18]

Finally, combining field and market data can certainly provide promising new insights into the conditions and mechanisms underlying discrimination, or also other social interactions that are embedded in market structures. In this article we focused on methodological issues, but one could use similar combinations of field experiments with market data also to study substantive questions of interest in (discrimination) research, such as how the tightness of markets or market power of different large suppliers affect their (discriminating) actions (for exemplary discussion on these issues see e.g. Ashenfelter and Hannan, 1986; Baert et al., 2013; Carlsson et al., 2018). We hope that our case study was also stimulating in that way.

18 However, this would also bring along some ethical issues. Testing the same supplier multiple times puts a higher, possibly disproportionately high, burden on this supplier. On the other hand, such an approach could provide unique insights into the mechanisms underlying discrimination. Ethics committees might decide what is reasonable.

Acknowledgments

For helpful suggestions, we thank participants of the conference "Analytical Sociology" at Venice International University in 2017. We are grateful for comments on earlier versions we received from two anonymous reviewers and from the editors. Maximilian Sonnauer helped us in compiling the database for the field experiments.

Data Note

We used data collected in the research project "Ethnic Discrimination and Segregation in German Housing Markets" funded by a small non-profit foundation in Germany, the Wolfgang and Anita Bürkle foundation. Replication files (Stata do-files and the field experimental data) can be found in the Supplemental Material accompanying this article.

References

Acolin, A., R. Bostic, and G. Painter. 2016. "A Field Study of Rental Market Discrimination Across Origins in France." *Journal of Urban Economics* 95:49–63.

Ahmed, Ali M., Lina Andersson, and Mats Hammarstedt. 2010. "Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants?" *Land Economics* 86:79–90.

Aigner, Dennis J. and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review* 30:175–187.

Andersson, Lisa R., Niklas Jakobsson, and Andreas Kotsadam. 2012. "A Field Experiment of Discrimination in the Norwegian Housing Market: Gender, Class, and Ethnicity." *Land Economics* 88:233–240.

Arrow, Kenneth J. 1973. "The Theory of Discrimination", pp. 3–33. In: Ashenfelter, Orley and Albert Rees (eds.): *Discrimination in Labor Markets*. Princeton: University Press.

Arrow, Kenneth J. 1998. "What Has Economics to Say About Racial Discrimination?" *Journal of Economic Perspectives* 12:91–100.

Ashenfelter, Orley and Timothy Hannan. 1986. "Sex Discrimination and Product Market Competition: The Case of the Banking Industry." *The Quarterly Journal of Economics* 101:149–173.

Auspurg, Katrin, Josef Brüderl, and Thomas Wöhler. 2019a. "Does Immigration Reduce the Support for Welfare Spending? A Cautionary Tale on Spatial Panel Data Analysis." *American Sociological Review* 84:754–763.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, California: Sage.

Auspurg, Katrin, Thomas Hinz, and Laura Schmid. 2017. "Contexts and Conditions of Ethnic Discrimination: Evidence from a Field Experiment in a German Housing Market." *Journal of Housing Economics* 35:26–36.

Auspurg, Katrin, Andreas Schneck, and Thomas Hinz. 2019b. "Closed Doors Everywhere? A Meta-Analysis of Field Experiments on Ethnic Discrimination in Rental Housing Markets." *Journal of Ethnic and Migration Studies* 45:95–114.

Baert, Stijn, Bart Cockx, Niels Gheyle, and Cora Vandamme. 2013. "Do Employers Discriminate Less If Vacancies Are Difficult to Fill? Evidence from a Field Experiment." Institute for the Study of Labor IZA Discussion Paper No. 7145.

Baldini, Massimo and Marta Federici. 2011. "Ethnic Discrimination in the Italian Rental Housing Market." *Journal of Housing Economics* 20:1–14.

Ball, Michael. 2016. "Housing Provision in 21st Century Europe." *Habitat International* 54:182–188.

Becker, Gary S. 1957. *The Economics of Discrimination.* Studies in Economics of the Economics Research Center of the University of Chicago. Chicago: University of Chicago Press.

Bell, Stephen H. and Elizabeth A. Stuart. 2016. "On the "Where" of Social Experiments: The Nature and Extent of the Generalizability Problem." *New Directions for Evaluation* 2016:47–59.

Bengtsson, Ragnar, Ellis Iverman, and Björn Tyrefors Hinnerich. 2012. "Gender and Ethnic Discrimination in the Rental Housing Market." *Applied Economic Letters* 19:1–5.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95:94–98.

Bertrand, Marianne and Esther Duflo. 2017. "Field Experiments on Discrimination", book section 8, pp. 309–393. In: Banerjee, Abhijit Vinayak and Esther Duflo (eds): *Handbook of Economic Field Experiments.* Amsterdam: North-Holland.

Biddle, Jeff E. and Daniel S. Hamermesh. 2013. "Wage Discrimination over the Business Cycle." *IZA Journal of Labor Policy* 2:1–19.

Björnsson, Davíð F., Fredrik Kopsch, and Gylfi Zoega. 2018. "Discrimination in the Housing Market as an Impediment to European Labour Force Integration: the Case of Iceland." *Journal of International Migration and Integration* 19:829–847.

Bosch, Mariano, M. Angeles Carnero, and Lídia Farré. 2010. "Information and Discrimination in the Rental Housing Market: Evidence from a Field Experiment." *Regional Science and Urban Economics* 40:11–19.

Bosch, Mariano, M. Angeles Carnero, and Lídia Farré. 2015. "Rental Housing Discrimination and the Persistence of Ethnic Enclaves." *SERIEs* 6:129–152.

Brewer, Marilynn B. and William D. Crano. 2014. "Research Design and Issues of Validity", pp. 11–26. In: Judd, Charles M. and Harry T. Reis (eds.): *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press.

Bunel, Mathieu, Samuel Gorohouna, Yannivk L'Horty, Pascale Petit, and Catherine Ris. 2019. "Ethnic Discrimination in the Rental Housing Market: An Experiment in New Caledonia." *International Regional Science Review* 42:65–97.

Carlsson, Magnus and Stefan Eriksson. 2014. "Discrimination in the Rental Market for Apartments." *Journal of Housing Economics* 23:41–54.

Carlsson, Magnus, Luca Fumarco, and Dan-Olof Rooth. 2018. "Does Labor Market Tightness Affect Ethnic Discrimination in Hiring?" Institute for the Study of Labor IZA Discussion Paper No. 11285.

Charness, Gary, Uri Gneezy, and Michael Kuhn. 2012. "Experimental Methods: Between-Subject and Within-Subject Design." *Journal of Economic Behavior & Organization* 81:1–8.

Cleveland, William S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74:829–836.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Drever, Anita I. and William A.V. Clark. 2002. "Gaining Access to Housing in Germany: The Foreign-minority Experience." *Urban Studies* 39:2439–2453.

Elwert, Felix and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31–53.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *Review of Economics and Statistics* 96:119–134.

Fernandez, Roberto M. and Santiago Campero. 2014. "Does Competition Drive Out Discrimination?" Paper Presented at the Annual Meeting of the American Sociological Association.

Flage, Alexandre. 2018. "Ethnic and Gender Discrimination in the Rental Housing Market: Evidence from a Meta-Analysis of Correspondence Tests, 2006-2017." *Journal of Housing Economics* 41:251–273.

Galster, George C. 1992. "Research on Discrimination in Housing and Mortgage Markets: Assessment and Future Directions." *Housing Policy Debate* 3:637–683.

Galster, George C. 1996. "Future Directions in Mortage Discrimination Research and Enforcement", pp. 697–716. In: Goering, John M. and Ron Wienk (eds.): *Mortgage Lending, Racial Discrimination, and Federal Policy*. Washington, DC: The Urban Institute Press.

German Statistical Office. 2017. *Nationalities in Germany (Table 12411-0009)*.

Greene, William H. 2012. *Econometric Analysis*. Boston: Prentice Hall.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology* 85:197–216.

Hanson, Andrew, Z. Hawley, and A. Taylor. 2011. "Subtle Discrimination in the Rental Housing Market: Evidence from E-Mail Correspondence with Landlords." *Journal of Housing Economics* 20:276–284.

Hanson, Andrew and Michael Santas. 2014. "Field Experiment Tests for Discrimination against Hispanics in the US Rental Housing Market." *Southern Economic Journal* 81:135–167.

Harrison, Malcolm, Ian Law, and Deborah Phillips. 2005. "Migrants, Minorities and Housing: Exclusion, Discrimination and Anti-Discrimination in 15 Member States of the European Union." Report, European Monitoring Centre on Racism and Xenophobia.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12:101–116.

Hellerstein, Judith, David Neumark, and Kenneth Troske. 1998. "Market Forces and Sex Discrimination." *The Journal of Human Resources* 37:353–380.

Heylen, Kristof and Katleen Van den Broeck. 2016. "Discrimination and Selection in the Belgian Private Rental Market." *Housing Studies* 31:223–236.

Hogan, Bernie and Brent Berry. 2011. "Racial and Ethnic Biases in Rental Housing: An Audit Study of Online Apartment Listings." *City and Community* 10:351–372.

Jann, Ben. 2014. "Plotting Regression Coefficients and Other Estimates." *Stata Journal* 14:708–737.

Krysan, Maria and Kyle Crowder. 2017. *The Cycle of Segregation: Social Processes and Residential Stratification*. New York: Russell Sage Foundation.

Lavallée, Pierre and Jean-François Beaumont. 2015. "Why We Should Put Some Weight on Weights." *Survey Methods: Insights from the Field* .

Mazziotta, Agostino, Michael Zerr, and Anette Rohmann. 2015. "The Effects of Multiple Stigmas on Discrimination in the German Housing Market." *Social Psychology* 46:325–334.

Metcalf, Gabriel. 2018. "Sand Castles Before the Tide? Affordable Housing in Expensive Cities." *Journal of Economic Perspectives* 32:59–80.

Murchie, Judson and Jindong D. Pang. 2018. "Rental Housing Discrimination Across Protected Classes: Evidence from a Randomized Experiment." *Regional Science and Urban Economics* 73:170–179.

Neumark, David. 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47:1128–1157.

Norton, Edward C, Hua Wang, and Chunrong Ai. 2004. "Computing interaction effects and standard errors in logit and probit models." *The Stata Journal* 4:154–167.

Oblom, A. and J. Antfolk. 2017. "Ethnic and Gender Discrimination in the Private Rental Housing Market in Finland: A Field Experiment." *Plos One* 12:e0183344.

Pager, Devah. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual review of sociology* 34:181–209.

Pager, Devah and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34:181.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62:659–661.

Quillian, Lincoln. 2006. "New Approaches to Understanding Racial Prejudice and Discrimination." *Annual Review of Sociology* 32:299–328.

Riach, Peter A. and Judith Rich. 2004. "Deceptive Field Experiments of Discrimination: Are they Ethical?" *Kyklos* 57:457–470.

Ross, Stephen and Margery Austin Turner. 2005. "Housing Discrimination in Metropolitan America: Explaining Changes between 1989 and 2000." *Social Problems* 52:152–180.

Salganik, Matthew. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, US: Houghton, Mifflin and Company.

StataCorp. 2015. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP.

Turner, Margery A, Stephen Ross, George C Galster, and John Yinger. 2002. *Discrimination in Metropolitan Housing Markets: National Results from Phase 1 of the Housing Discrimination Study (HDS)*. Washington, DC: Urban Institute.

van Es, Bert, Chris A. J. Klaassen, and Karin Oudshoorn. 2000. "Survival Analysis Under Cross-Sectional Sampling: Length Bias and Multiplicative Censoring." *Journal of Statistical Planning and Inference* 91:295–312.

Vuolo, Mike, Christopher Uggen, and Sarah Lageson. 2016. "Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests With Nominal Outcomes." *Sociological Methods & Research* 45:260–303.

Wienk, Ronald E., Clifford E. Reid, John C. Simonson, and Frederick J. Eggers. 1979. *Measuring Racial Discrimination in American Housing Markets - The Housing Markets Practices Survey*. Washington, DC: U.S. Department of Housing and Urban Development.

Yinger, John. 1986. "Measuring Racial Discrimination with with Fair Housing Audits: Caught in the Act." *American Economic Review* 76:881–893.

4.A. Appendix

4.A.1. Monte Carlo Simulations

Table 4.A1 shows the results of Monte Carlo simulations of different sampling strategies. For each sampling strategy, 50 random samples were drawn out of the market data that were collected during our 1 year observation period. The simulations are conducted for sampling intervals of one week (column 1), one month (column 2) and six months (column 3). For each interval, random samples were drawn once on the apartment level (with replacement of suppliers) and once on supplier level (without replacement of suppliers). The whole market serves as a reference category and is described in column 4.

As can be seen, the sizes of suppliers in the drawn samples on the apartment level do not differ drastically from the sizes of suppliers we observe in the whole market. Sampling on supplier level, on the contrary, leads to a systematic over-representation of small suppliers—as one would expect.

For the advertisement time, we find the sampled apartments to be considerably longer advertised than we observe for the whole market. The longer the sampling interval, both for sampling on supplier level and sampling on apartment level, the shorter the advertisement time in the sample. Thus, with longer sampling intervals the gap to the observed advertisement times decreases, as one would expect. However, even for the sampling interval of six months, there remains a rather large difference to the observed advertisement times observed for the whole market.

To understand why this rather large gap between the sampled and observed advertisement times remains even for the sampling interval of six months the used sampling strategy has to be discussed in more detail. To mirror sampling strategies used by real apartment seekers, we used a *prospective* sampling approach. On each day in each sampling period, an equal fraction of advertisements ($N_{\text{sample}}/N_{\text{days}}$) was sampled, until the intended sample size was reached. Due to sampling sequentially on each day in the field period (instead of drawing one sample for the whole field period), a length bias remains also with long sampling periods: Each day, units with longer advertisement times had a higher chance of being sampled than those with shorter advertisement times. Only when drawing *one* joint sample for the whole observation period, this sampling bias would disappear.

Thus, for apartments being advertised for longer times using a prospective strategy

Table 4.A1.: Monte Carlo Simulations for Different Sampling Strategies

	(1) One week		(2) One month		(3) Six months		(4) Market data
	Mean	SD	Mean	SD	Mean	SD	Mean
Sampling on apartment level:							
Advertisement time							
Mean[a]	87.74	1.30	88.14	1.20	82.38	1.31	22.19
Median[a]	60.95	1.57	60.50	1.37	58.06	1.25	13.00
Size of supplier							
Mean	1,243.02	49.60	1,286.79	65.31	1,324.49	57.25	1,246.13
Median	36.72	1.59	37.92	1.94	41.59	1.70	34.00
Observations	50		50		50		695,458
Sampling on supplier level:							
Advertisement time							
Mean[a]	70.18	3.11	70.90	4.65	59.85	5.16	22.19
Median[a]	43.99	3.60	44.20	4.71	42.33	3.44	13.00
Size of supplier							
Mean	21.48	10.20	27.42	22.89	35.86	70.61	1,246.13
Median	1.00	0.00	1.00	0.00	1.27	0.65	34.00
Observations	50		50		50		695,458

Notes: The sampling period for all simulations ended on 31st October 2015, and started the indicated time before: For column (1) on 25th October, (2) on 1st October, and (3) on 1st May 2015. For comparison, descriptive statistics on the whole market data are given (4). [a]Due to censored information on the advertisement time, we can calculate this statistic only based on 575,950 cases in the market data (see also Table 4.2 in the main text).

still results in a higher probability of being sampled.[19] However, only the prospective sampling likely matches the search strategies of real apartment seekers; and to our knowledge, this technique is also common in research so far. Using an alternative retrospective sample, i.e. drawing a sample of units that were advertised during the last week, month or even half year, there would be a very high risk that apartments are no longer available when the experiment takes place. For similar reasons, the market data used in our study would not provide an adequate sampling frame. For the length bias that exists between field experiments in contrast to real apartment seekers, therefore the comparison of different samples (and not the contrast to the market data) is of most interest. (The whole advertisement time found in the market data might, however, be used in future research as an indicator for the tightness of markets or pickiness of landlords.)

4.A.2. Regression Results

19 Therefore, we also conducted a *retrospective* approach. With this approach, with increasing sampling periods the advertisement times approximate those in the market data at a much faster rate than in the prospective apporach. However, using the retrospective approach also strengthens the over-representation of small suppliers with longer sampling periods when sampling on the supplier level.

Table 4.A2.: Multinomial Logistic Regression Models of Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant'; Logit Coefficients and p-Values (in Parentheses)

	M1 Only Controls		M2 + Apartment Characteristics		M3[a] + Applicant Characteristics		M4a[a,b] + Interaction w/ Supplier Size		M4b[a,c] + Interaction w/ Advertisement Time	
	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G
Percentage of foreigners	-0.2893^{+}	0.1027	-0.2949^{+}	0.1186	-0.2954^{+}	0.1434	-0.2898^{+}	0.1395	-0.3036^{+}	0.1611
	(0.0951)	(0.6681)	(0.0918)	(0.6198)	(0.0867)	(0.5616)	(0.0839)	(0.5772)	(0.0846)	(0.5140)
City	0.0151	−0.2235	0.0334	−0.2183	0.0434	−0.2510	0.0417	−0.2470	0.0549	−0.2443
	(0.9225)	(0.3839)	(0.8285)	(0.3977)	(0.7771)	(0.3318)	(0.7825)	(0.3438)	(0.7220)	(0.3554)
Log size of supplier			-0.0830^{**}	−0.1128	-0.0809^{**}	−0.1111	−0.0121	−0.3330	-0.0824^{**}	-0.1230^{+}
			(0.0077)	(0.1139)	(0.0095)	(0.1289)	(0.9170)	(0.1461)	(0.0092)	(0.0955)
Log advertisement time			−0.0519	0.0525	−0.0516	0.0556	−0.0500	0.0615	0.3875^{*}	0.3663
			(0.2839)	(0.5389)	(0.2935)	(0.5257)	(0.3073)	(0.4686)	(0.0326)	(0.1597)
Applicant characteristics										
Educational status										
(ref.: high)										
No info					0.2264	0.2293	0.1475	−0.0548	0.3645	0.5660
					(0.1230)	(0.3085)	(0.3609)	(0.8411)	(0.3457)	(0.3479)
Low					0.2409^{+}	0.0445	0.1863	−0.0386	0.8573^{*}	0.3227
					(0.0653)	(0.8592)	(0.2405)	(0.8986)	(0.0148)	(0.5807)
Employment status										
(ref.: employed)										
No info					0.1678	0.1410	0.2721	0.0508	0.5981	1.4620^{+}
					(0.2787)	(0.5663)	(0.1820)	(0.8705)	(0.1107)	(0.0697)

(continued on next page)

(continued)

	M1		M2		M3[a]		M4a[a,b]		M4b[a,c]	
	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G
Self-employed					0.2335^{+}	0.0408	0.2217	−0.0805	0.1626	1.3064
					(0.0747)	(0.8692)	(0.1974)	(0.8069)	(0.7343)	(0.1484)
Public employee					−0.0022	−0.2674	−0.0868	−0.1913	0.3614	1.1642
					(0.9892)	(0.3415)	(0.6643)	(0.5834)	(0.3396)	(0.1739)
Interaction: log size of supplier										
× Applicant characteristics										
Educational status										
(ref.: high)										
No info							0.0766	0.3020*		
							(0.3414)	(0.0328)		
Low							0.0512	0.1221		
							0.0766	0.3020*		
Employment status										
(ref.: employed)										
No info							−0.1122	0.1030		
							(0.3618)	(0.5818)		
Self-employed							0.0038	0.1065		
							(0.9689)	(0.5739)		
Public employee							0.0688	−0.0488		
							(0.4894)	(0.8054)		
Interaction: log advertisement time										
× Applicant characteristics										
Educational status										
(ref.: high)										

(continued on next page)

(continued)

	M1		M2		M3[a]		M4a[a,b]		M4b[a,c]	
	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G
No info									−0.0532	−0.1159
									(0.6624)	(0.5437)
Low									−0.2232^{+}	−0.0915
									(0.0559)	(0.6235)
Employment status (ref.: employed)										
No info									−0.1574	−0.4352^{+}
									(0.2184)	(0.0704)
Self-employed									0.0256	−0.4198
									(0.8736)	(0.1274)
Public employee									−0.1329	−0.4780^{+}
									(0.2842)	(0.0701)
Constant	−1.3750***	−3.1053***	−1.1255***	−3.1476***	−1.2380***	−3.3177***	−1.3427***	−3.2050***	−2.4843***	−4.3762***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)
N	2,992		2,992		2,992		2,992		2,992	
AIC	3,399.86		3,396.48		3,416.91		3,436.39		3,435.32	
BIC	3,435.88		3,456.52		3,645.05		3,832.63		3,831.56	

Notes: p-values are in parentheses; $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. For outcome 'Discr. T' ('Discr. G'), applicant's characteristics are those of Turks (Germans). [a]Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier. [b]Also controlled for the interaction of log size of supplier and applicant's family status as well as all characteristics of the applicant of the other ethnicity. [c]Also controlled for the interaction of log advertisement time and applicant's family status as well as all characteristics of the applicant of the other ethnicity.

Table 4.A3.: Multinomial Logistic Regressions, AMEs of Predictors and p-Values (in Parentheses) for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment') as Reported in Figure 4.5 (in the Main Text)

	Discr. T	Discr. G
Percentage of foreigners	-0.0368^{+}	0.0069
	(0.0825)	(0.4610)
City	0.0067	-0.0094
	(0.7194)	(0.3234)
Log size of supplier	-0.0092^{*}	-0.0036
	(0.0157)	(0.1864)
Log advertisement time	-0.0066	0.0023
	(0.2716)	(0.4713)
Applicant characteristics		
Educational status (ref.: high)		
No info	0.0277	0.0087
	(0.1129)	(0.2929)
Low	0.0287^{+}	0.0017
	(0.0590)	(0.8425)
Employment status (ref.: employed)		
No info	0.0181	0.0059
	(0.3323)	(0.5370)
Self-employed	0.0265^{+}	0.0022
	(0.0957)	(0.8103)
Public employee	-0.0020	-0.0092
	(0.9132)	(0.3046)
N	2,992	
AIC	3,416.91	
BIC	3,645.05	

Notes: p-values are in parentheses; $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. For outcome 'Discr. T' ('Discr. G'), applicant's characteristics are those of Turks (Germans). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.

Table 4.A4.: Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with p-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as Reported in Figure 4.6 (in the Main Text)

	Supplier Size			Advertisement Time		
	small	medium	large	short	medium	long
Percentage of foreigners	−0.0232	−0.0225	−0.0740*	−0.0317	−0.0480^{+}	−0.0224
	(0.4232)	(0.3816)	(0.0210)	(0.2585)	(0.0917)	(0.5579)
City	−0.0185	0.0118	0.0531^{+}	−0.0302	0.0375	0.0030
	(0.4809)	(0.7175)	(0.0867)	(0.3080)	(0.2148)	(0.9268)
Applicant characteristics						
Educational status (ref.: high)						
No info	0.0086	0.0574^{+}	0.0458	0.0326	0.0454^{+}	0.0025
	(0.6933)	(0.0578)	(0.1607)	(0.2099)	(0.0946)	(0.9349)
Low	0.0216	0.0286	0.0369	0.0774**	0.0252	−0.0228
	(0.3495)	(0.4173)	(0.2039)	(0.0010)	(0.3390)	(0.4348)
Employment status (ref.: employed)						
No info	0.0284	0.0235	−0.0101	0.0681*	−0.0146	−0.0027
	(0.3229)	(0.4938)	(0.7911)	(0.0115)	(0.6721)	(0.9326)
Self-employed	0.0281	0.0453	0.0137	0.0742**	−0.0118	0.0164
	(0.2488)	(0.1866)	(0.7034)	(0.0099)	(0.7199)	(0.5855)
Public employee	−0.0232	0.0414	0.0215	0.0349	−0.0281	−0.0248
	(0.3350)	(0.2504)	(0.5362)	(0.1978)	(0.3985)	(0.4066)
N	1,661	658	673	1,044	980	968
AIC	2,057.62	718.66	708.55	1,210.51	1,162.51	1,105.55
BIC	2,241.73	871.29	861.95	1,378.84	1,328.68	1,271.30

Notes: p-values are in parentheses; $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment. Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.

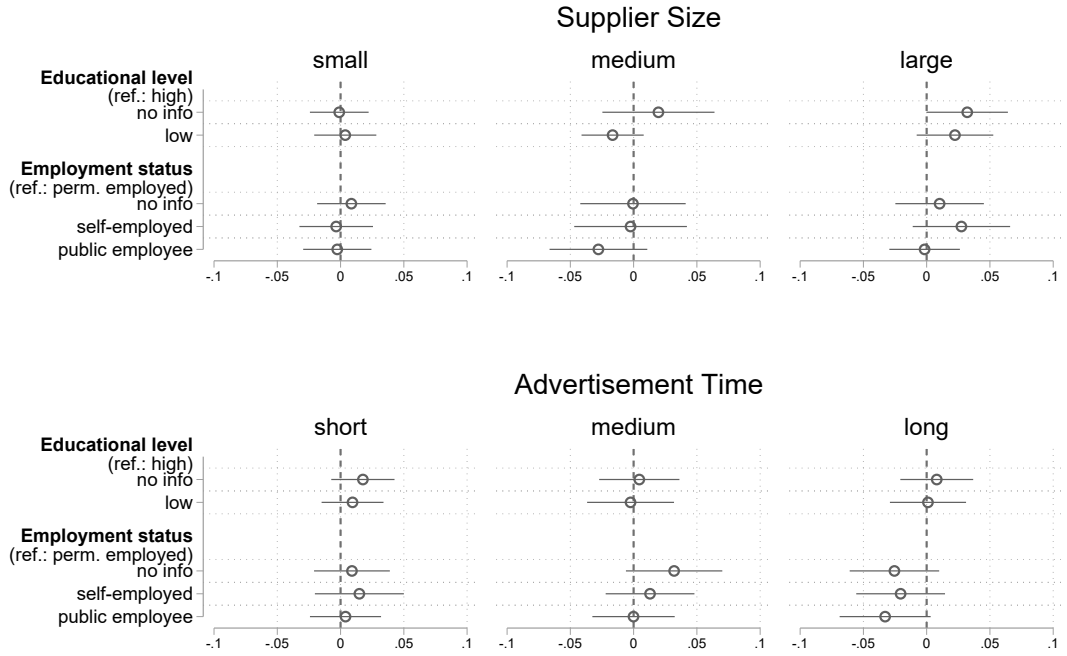

Figure 4.A1.: Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time

Notes: Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); (top panel); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment (below panel). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table 4.A5 in the Appendix for the AMEs displayed. All estimates are based on at least $N = 658$ observations.

Table 4.A5.: Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with p-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as reported in Figure 4.A1 (in the Appendix)

	Supplier Size			Advertisement Time		
	small	medium	large	short	medium	long
Percentage of foreigners	0.0041	0.0153	0.0086	0.0142	0.0006	0.0036
	(0.7333)	(0.2748)	(0.6287)	(0.2374)	(0.9697)	(0.8156)
City	−0.0031	−0.0276	−0.0077	−0.0087	−0.0329^{+}	0.0126
	(0.8281)	(0.1581)	(0.6946)	(0.6257)	(0.0533)	(0.4716)
Applicant characteristics						
Educational status (ref.: high)						
No info	−0.0010	0.0196	0.0321^{+}	0.0177	0.0045	0.0080
	(0.9314)	(0.3856)	(0.0503)	(0.1663)	(0.7809)	(0.5878)
Low	0.0037	−0.0166	0.0224	0.0095	−0.0024	0.0010
	(0.7671)	(0.1852)	(0.1473)	(0.4480)	(0.8904)	(0.9461)
Employment status (ref.: employed)						
No info	0.0086	−0.0005	0.0102	0.0090	0.0320^{+}	−0.0255
	(0.5354)	(0.9809)	(0.5662)	(0.5555)	(0.0994)	(0.1578)
Self-employed	−0.0034	−0.0024	0.0275	0.0149	0.0130	−0.0206
	(0.8165)	(0.9159)	(0.1621)	(0.4072)	(0.4678)	(0.2502)
Public employee	−0.0026	−0.0278	−0.0016	0.0039	−0.0001	−0.0328^{+}
	(0.8504)	(0.1582)	(0.9088)	(0.7849)	(0.9970)	(0.0738)
N	1,661	658	673	1,044	980	968
AIC	2,057.62	718.66	708.55	1,210.51	1,162.51	1,105.55
BIC	2,241.73	871.29	861.95	1,378.84	1,328.68	1,271.30

Notes: p-values are in parentheses; $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment. Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.