Fehr-Duda, Helga; Schimmelpfennig, Robin

Working Paper

Wider die Zahlengläubigkeit: Sind Befragungsergebnisse eine gute Grundlage für wirtschaftspolitische Entscheidungen?

Working Paper, No. 297

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Fehr-Duda, Helga; Schimmelpfennig, Robin (2018) : Wider die Zahlengläubigkeit: Sind Befragungsergebnisse eine gute Grundlage für wirtschaftspolitische Entscheidungen?, Working Paper, No. 297, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-153525

This Version is available at:
https://hdl.handle.net/10419/192906

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 297

Wider die Zahlengläubigkeit: Sind Befragungsergebnisse eine gute Grundlage für wirtschaftspolitische Entscheidungen?

Helga Fehr-Duda und Robin Schimmelpfennig

Revised version, December 2018

11. Dezember 2018

Helga Fehr-Duda* und Robin Schimmelpfennig

Wider die Zahlengläubigkeit: Sind Befragungsergebnisse eine gute Grundlage für wirtschaftspolitische Entscheidungen?

Zusammenfassung: Befragungen von Konsumenten, Steuerzahlern und Wählern werden in Wirtschaft und Politik häufig als Entscheidungsgrundlage herangezogen. Die Ergebnisse dieser Befragungen können somit großen Einfluss auf politische Entscheidungen und damit auf den Alltag vieler Menschen haben. Befragungen sind besonders dann wichtig und notwendig, wenn noch keine entsprechenden Verhaltensdaten oder Erfahrungswerte vorliegen. Helga Fehr-Duda und Robin Schimmelpfennig gehen der Frage nach, ob Angaben zur Zahlungsbereitschaft, die in hypothetischen Entscheidungssituationen gemacht werden, glaubwürdig und zuverlässig genug sind, um als Grundlage für unternehmerische und wirtschaftspolitische Entscheidungen zu dienen. Die Autoren zeigen anhand zweier neuer Studien, dass Ergebnisse hypothetischer Befragungen, wie beispielsweise einer Stated preference Befragung, signifikant vom Kontext der Entscheidungssituation abhängen. Es ist daher nicht möglich, allgemeine kontextunabhängige Aussagen zum Ausmaß der möglichen Fehleinschätzung von Zahlungsbereitschaften zu treffen. Die Autoren empfehlen, wann immer möglich Labor- oder Feldexperimente im relevanten Kontext mit realen monetären Anreizen durchzuführen. In jenen Fällen, in denen hypothetische Befragungen die einzige Möglichkeit der Datengewinnung darstellen, ist es unabdingbar, das Befragungsdesign möglichst anreizverträglich, realitätsnah und dem Anwendungskontext angemessen zu gestalten. Auftraggeber sollten auf die Einhaltung dieser Grundsätze bestehen, um möglichst valide Befragungsergebnisse zu erhalten.

JEL-Klassifikation: A12, B41, D03

Schlüsselwörter: Verhaltensökonomie, experimentelle Wirtschaftsforschung, Stated preference Befragung, Kontingente Bewertung, Zahlungsbereitschaft

***Kontaktperson: Helga Fehr-Duda**, Institut für Banking und Finance, Universität Zürich, Plattenstrasse 32, 8032 Zürich, Schweiz, E – Mail: helga.fehr@bf.uzh.ch
Robin Schimmelpfennig, Department of Psychological and Behavioural Science, London School of Economics and Political Science, 55/56 Lincoln’s Inn Fields, London WC2A 3LJ, Vereinigtes Königreich

Inhaltsverzeichnis

1 Der Diskurs um die Validität von Befragungsergebnissen

Viele Leser können sich vielleicht noch an das Jahr 1989 erinnern, als verstörende Bilder von Tausenden ölverschmutzter Seevögel durch die Medien gingen. Im Prinz-William-Sund vor der Küste Alaskas war der Supertanker Exxon Valdez auf Grund gelaufen und hatte eine Ölpest ungeheuren Ausmaßes ausgelöst. Dieses Ereignis hatte nicht nur massive Auswirkungen auf die Umwelt, sondern auch auf die amerikanische Rechtsprechung. Zum ersten Mal in deren Geschichte war nicht nur der direkte Schaden für Wirtschaft und Umwelt Gegenstand der Beurteilung der Schadenshöhe, sondern auch der indirekte Schaden durch Verlust von sogenanntem „Existenzwert“ (Existence value), auch passiver Nutzwert (Passive use value) genannt (Krutilla 1967). Mit diesem Begriff meint man jenen Nutzen, der aus dem bloßen Wissen über die Existenz von Umweltgütern resultiert, wie z.B. dem Fortbestand von bedrohten Tierarten oder sauberen Küstenabschnitten, selbst wenn man diese Orte nie aufsucht. Im Gefolge der rechtlichen Aufarbeitung des Tankerunfalls wurde im sogenannten Oil Pollution Act 1990 die National Oceanic und Atmospheric Administration (NOAA) mit der Ausarbeitung von Richtlinien zur Ermittlung des Gesamtschadens inklusive des Verlusts an Existenzwert beauftragt (Pierce 1994).

Die Schwierigkeit bei der Berücksichtigung von Existenzwerten liegt darin, dass diese nicht auf Märkten gehandelt werden und daher keine Marktpreise, das übliche ökonomische Maß des Wertes eines bestimmten Gutes, beobachtbar sind. Mit der Exxon Valdez Ölpest ist daher das Verfahren der „Kontingenten Bewertung” (Contingent valuation) ins Zentrum der Aufmerksamkeit gerückt, das verspricht, den Wert nicht-handelbarer Güter mittels Fragebögen oder Interviews zu ermitteln. Die Methode der Kontingenten Bewertung beruht also nicht auf tatsächlichen Verhaltensdaten, sondern auf Angaben in hypothetischen Entscheidungssituationen, auf sogenannten „geäußerten Präferenzen“ (Stated preferences). Stated preference Befragungen (SP-Befragungen) werden in unterschiedlichen Formaten durchgeführt und vor allem dann eingesetzt, wenn es gilt, die Nachfrage nach Gütern und Dienstleistungen abzuschätzen, die entweder noch nicht auf Märkten gehandelt werden oder die prinzipiell nicht handelbar sind. Sie kommen daher nicht nur speziell für die Messung von Existenzwerten, sondern auch ganz allgemein für die Bewertung von öffentlichen und privaten Gütern in der Gesundheitsökonomie, Agrarökonomie, im Marketing und in der Mobilitätsforschung zur Anwendung. Aber können Zahlen, die mittels hypothetischer Fragen erhoben werden, verlässlich die tatsächliche Bewertung eines Gutes abbilden?

Um dieser Frage nachzugehen, berief die NOAA 1992 eine Arbeitsgruppe unter der Leitung der Nobelpreisträger Kenneth Arrow und Robert Solow ein. Diese Arbeitsgruppe hatte den Auftrag zu untersuchen, ob Antworten auf hypothetische Fragen verlässlich genug seien, um Schäden inklusive des Verlusts von Existenzwerten abzuschätzen und damit als Grundlage für die Rechtsprechung dienen zu können. Die NOAA-Arbeitsgruppe kam zu einem vorsichtig bejahenden Befund und veröffentlichte 1994 gemäß Oil Pollution Act Richtlinien zur Durchführung von solchen Befragungen (Portney 1994).

Die akademische Debatte zu dieser Zeit verlief indessen äußerst kontrovers. In einem berühmten Artikel mit dem Titel „Contingent valuation: Is some number better than no number?“ geißelten Diamond und Hausman (1994) die Methode der Kontingenten Bewertung wegen ihrer mangelnden Glaubwürdigkeit und Verlässlichkeit.[1] Einer ihrer Hauptkritikpunkte bezog sich auf die zu schwache Reaktion von gemessenen Zahlungsbereitschaften auf Mengenänderungen: Beispielsweise unterscheidet sich die erhobene Zahlungsbereitschaft für die Reinigung eines einzigen verschmutzten Sees kaum von jener für die Reinigung vieler verschmutzter Seen (Scope effect, Embedding effect).

In der Zwischenzeit wurden Tausende von SP-Befragungen in den verschiedensten Bereichen durchgeführt und große Anstrengungen unternommen, Befragungsdesign und ökonometrische Auswertungsverfahren zu verbessern. Trotzdem kommt bald 20 Jahre nach seiner ursprünglichen skeptischen Einschätzung Hausman (2012) zu einem vernichtenden Verdikt: Kontingente Bewertung sei nicht nur eine zweifelhafte Methode, sondern sogar hoffnungslos, sodass man sie nicht als Grundlage für wirtschaftspolitische Entscheidungen oder zur Schadensfestsetzung in Gerichtsverfahren verwenden sollte. Neben dem schon erwähnten Problem der mangelnden Mengensensitivität führt Hausman (2012) vor allem Hypothetical bias als Begründung seines Urteils an. Unter Hypothetical bias versteht man die Diskrepanz zwischen „Sagen“ und „Tun“, d.h. die in einer Befragung geäußerte Absicht, etwas Bestimmtes zu tun, beispielsweise ein bestimmtes Produkt zu einem bestimmten Preis zu kaufen, wird in einer realen Entscheidungssituation nicht umgesetzt. Für die Interpretation von gemessenen Zahlungsbereitschaften bedeutet Hypothetical bias, dass die erhobenen Werte mit einem Messfehler behaftet sind. Es wäre also äußerst wichtig zu wissen, in welche Richtung die Verzerrung – Überschätzung oder Unterschätzung – geht und wie groß das Ausmaß der Fehleinschätzung ist. Angaben zu Präferenzen und Zahlungsbereitschaft sind nur dann für Wirtschaftspolitik, Unternehmen und Rechtsprechung brauchbar, wenn sie extern valide sind, d.h. wenn sie tatsächliches Verhalten verlässlich prognostizieren können.

Louviere bringt die Anforderungen an eine gute Studie folgendermaßen auf den Punkt: „. . . the external validity [...] likely rests on the degree to which [it] simulates all key aspects of a real decision, including the incentive compatibility of questions, framing of situations, contexts, consequences, etc. If one can design experiments that simulate real choice situations as closely as possible, one should be more likely to obtain results that mimic real life. If one fails to capture all the relevant aspects of a real situation, there should be systematic deviations. It’s that simple. It’s also that complex“ (Louviere 2006, S. 174). Externe Validität hängt also entscheidend von der Realitätsnähe der Entscheidungssituationen ab, die unter anderem von Anreizverträglichkeit, dem Vorhandensein von realen Konsequenzen, dem Format der Fragestellung und dem Kontext der Entscheidungen beeinflusst wird. Anreizverträglichkeit bedeutet, dass die Fragen so gestaltet werden müssen, dass es

[1] Diamond erhielt 2010 gemeinsam mit Mortensen und Pissarides den wirtschaftswissenschaftlichen Nobelpreis.

im besten Interesse der Befragten ist, ihre Präferenzen wahrheitsgemäß zu offenbaren.

Im Folgenden stellen wir die üblichen SP-Befragungsformate vor, diskutieren ihre potentiellen Mängel im Hinblick auf Louvieres Kriterien und gehen insbesondere auf den Hypothetical bias ein. Während sich die Fachliteratur vergleichsweise intensiv mit dem Problem der Anreizverträglichkeit beschäftigt hat, legen wir hier den Fokus auf die Kontextabhängigkeit von Entscheidungen. Wir zeigen anhand einer Studie zur Verkehrsmittelwahl, wie wenig aussagekräftig Befragungsergebnisse sein können, wenn sie den relevanten Kontext vernachlässigen. Ein weiteres Beispiel demonstriert, welch dramatische Effekte die konkreten Umstände der Entscheidungssituation haben können. Abschließend versuchen wir, eine Antwort auf folgende Frage zu finden: Kann man Befragungsergebnissen wirklich nicht trauen oder gibt es Bedingungen, unter denen verlässliche Werte zu erwarten sind?

2 Methoden der Stated preference Befragung

Es gibt viele Methoden, mit deren Hilfe Daten zu wirtschaftspolitisch relevanten Entscheidungen gewonnen werden können. Dazu zählen unter anderem schriftliche Befragungen, mündliche Interviews, Laborexperimente, Feldexperimente und seit vielen Jahren auch webbasierte Experimente. Je nach Fragestellung können Experimente im Prinzip mit realen Anreizen versehen werden, indem die Entscheidungen der Experimentteilnehmenden tatsächliche monetäre Konsequenzen nach sich ziehen. D.h. die Teilnehmenden werden gemäß ihren Entscheidungen entlohnt, was einen nicht unerheblichen Kostenfaktor darstellen kann. SP-Befragungen hingegen sind immer hypothetischer Natur. Dabei kann eine fixe, entscheidungsunabhängige, Teilnahmeentschädigung zur Auszahlung kommen. Während diese verschiedenen Methoden hinsichtlich Kosten, Kapazität und Kontrollierbarkeit unterschiedlichen Beschränkungen unterliegen, werden SP-Befragungen von vielen Forschern als eine effektive Methode angesehen, um kostengünstig eine große Anzahl an Beobachtungen zu generieren. Im Wesentlichen kann man zwischen drei Kategorien von SP-Befragungen unterscheiden: Kontingente Bewertungsmethode, Choice Experiment und Conjoint Analyse, die sich je nach Einsatzgebiet in verschiedenen Details unterscheiden.

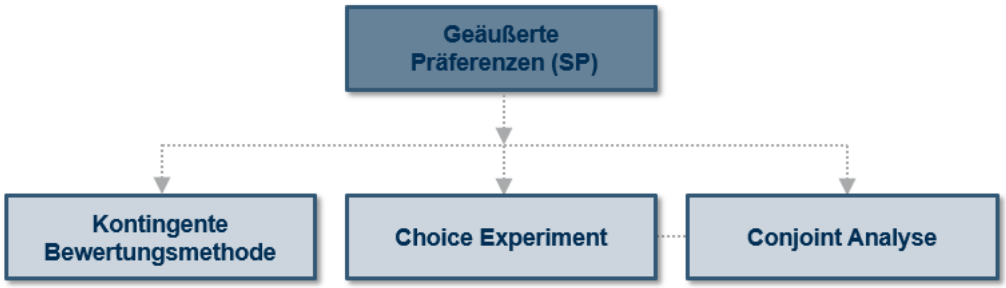

Abbildung 1: Übersicht der Befragungsmethoden für die Messung von geäußerten Präferenzen (Kling et al. 2012)

2.1 Kontingente Bewertung

Die durch den Untergang der Exxon Valdez 1989 verursachte Umweltkatastrophe dürfte unter anderem begründen, warum die Kontingente Bewertungsmethode die bekannteste SP-Befragungsmethode zur Ermittlung von Zahlungsbereitschaften (Willingness to pay – WTP) für Umweltgüter darstellt (Carson et al. 2003). Während es keine allgemein gültige Vorgehensweise gibt, basiert die Kontingente Bewertung doch auf wohldefinierten Elementen: Fixe Bestandteile sind einerseits eine genaue Beschreibung des zu bewertenden Projektes oder Programmes inklusive dessen Kosten, andererseits die Festlegung der Art der Präferenzmessung. Die NOAA-Richtlinien sehen beispielsweise vor, dass die Zahlungsbereitschaft für das Vermeiden zukünftiger Umweltschäden gemessen und die Präferenzen mittels Referendumsformat erhoben werden sollen. D.h. die Studienteilnehmenden müssen dem beschriebenen Projekt entweder zustimmen oder es ablehnen. Die Fragen werden also üblicherweise mit Ja/Nein beantwortet (Kling et al. 2012), es kommen aber auch andere Formate, wie offene Fragen, zum Einsatz. Der Anwendungsbereich der Kontingenten Bewertung ist nicht auf öffentliche Güter beschränkt, es gibt auch eine große Zahl von Studien zu privaten Gütern (z.B. Jagdlizenzen (Hammack und Brown 1974), Hochwasserversicherung (Botzen und van den Bergh 2012)).

2.2 Choice Experiment und Conjoint Analyse

Zwei weitere Unterkategorien der SP-Befragung bilden das Choice Experiment und die Conjoint Analyse, die vor allem für die Messung von Zahlungsbereitschaften im Marketing und in der Mobilitätsforschung zur Anwendung kommen. Oberflächlich betrachtet unterscheiden sich Choice Experiment und Conjoint Analyse in der Umsetzung kaum. Die Grundidee besteht darin, das zu bewertende Gut durch verschiedene Merkmale, wie z. B. Kaufpreis, Farbe, Stromverbrauch, Leistung etc., zu charakterisieren. Nehmen die Ausprägungen der Produktmerkmale diskrete Werte an, wird ein Choice Experiment als Discrete Choice Experiment bezeichnet. Die Ausprägungen dieser Merkmale werden systematisch variiert und zu Produktszenarien kombiniert. Die Befragungsteilnehmenden erhalten dann verschiedene Produktszenarien zur Auswahl, die jeweils auf einem

Entscheidungsblatt beschrieben werden. Üblicherweise werden die Teilnehmenden mit vielen verschiedenen Entscheidungsblättern konfrontiert, in denen sie zwischen den angegebenen Produktszenarien wählen müssen. Abbildung 2 zeigt ein typisches Entscheidungsblatt einer einfachen Entscheidungssituation aus der Verkehrsmittelwahl: Hier wird den Teilnehmenden die Wahl zwischen einer Fahrt mit dem Auto oder mit einem öffentlichen Verkehrsmittel vorgelegt, wobei das „Produkt" Fahrt durch die Produktmerkmale Reisezeit und Fahrtkosten in bestimmten Ausprägungen beschrieben wird. In diesem Beispiel betragen die Gesamtreisezeit mit dem Auto 20 Minuten und die Benzinkosten CHF 1,80. Durch Variation von Zeit und Kosten können mittels einer statistischen Auswertung die Teilnutzen der einzelnen Produktmerkmale berechnet werden.

In Choice Experimenten und Conjoint Analysen werden die Entscheidungssituationen in ähnlicher Weise generiert, die Methoden unterscheiden sich aber vor allem in ihrem theoretischen Hintergrund. In der Conjoint Analyse steht die statistische Effizienz des Experimentdesigns im Vordergrund, während die Auswahl der Produktmerkmale und ihrer Ausprägungen eher ad hoc erfolgt (Louviere et al. 2010). Choice Experimente sind der stochastischen Nutzentheorie (Random Utility Theory) verpflichtet, welche die inhärente Fehleranfälligkeit von Entscheidungen berücksichtigt und daher der Spezifikation des Entscheidungsfehlers große Bedeutung beimisst (Manski und McFadden 1981). Louviere et al. wehren sich gegen die Gleichsetzung von Conjoint Analyse und Choice Experiment folgendermaßen: „[Conjoint analysis] should be seen for what it really is, namely a purely descriptive way to fit a statistical model to a set of observed ranking or rating data with no ability to inform questions about how consumer behavior is likely to change in response to changes in choice context." (Louviere et al. 2010, S. 69).

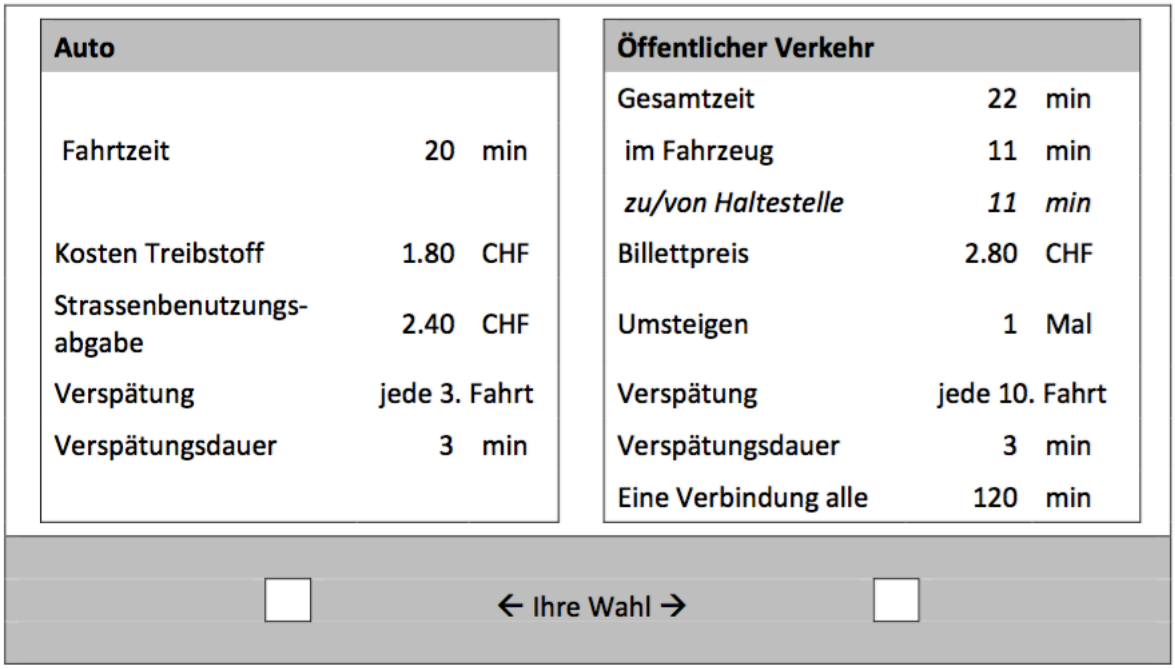

Abbildung 2: In dem Entscheidungsblatt werden die monetären und zeitlichen Kosten für eine Fahrt per Auto oder öffentlichem Verkehr gezeigt (nach Weis et al. 2016, S. 99)

2.3 Wann sind Umfrageergebnisse vertrauenswürdig und wann nicht?

SP-Befragungen werden häufig mit dem Ziel eingesetzt, Maßnahmen in Wirtschaft und Politik zu rechtfertigen. Jedoch muss die Frage, welche Diamond und Hausman (1994) schon vor über 20 Jahren gestellt haben, auch heute noch immer wieder aufs Neue beantwortet werden: „Is some number better than no number"? Befragungsergebnisse sind nur dann uneingeschränkt eine gute Grundlage für Entscheidungen, wenn die damit gewonnenen Daten valide sind. Mit anderen Worten: Werden die Bedingungen für realitätsnahe Entscheidungssituationen, die Louviere (2006) formuliert, nämlich Anreizverträglichkeit und das Vorhandensein von Konsequenzen, sowie die Art der Fragestellung und die Berücksichtigung des Kontexts, von SP-Befragungen erfüllt?

- Hypothetical bias und fehlende Anreize

Ein gewichtiger Kritikpunkt an SP-Befragungen besteht darin, dass die Entscheidungssituationen hypothetischer Natur sind, die keine spürbaren Konsequenzen nach sich ziehen. Carson und Groves (2007) identifizieren zwei Bedingungen, die notwendigerweise eingehalten werden müssen, damit die Befragungsteilnehmenden aufmerksam und wahrheitsgetreu antworten: Das Vorhandensein von Konsequenzen (Consequentiality) und die Anreizverträglichkeit (Incentive compatibility). Anreizverträglichkeit bedeutet, dass die Fragen so gestaltet werden müssen, dass es im besten Interesse der Befragten ist, ihre wahren Präferenzen zu offenbaren. Wenn Entscheidungen folgenlos bleiben, gibt es keinen Anreiz, die Wahrheit zu sagen. In der

experimentellen Wirtschaftsforschung werden daher, wann immer möglich, Entscheidungen an monetäre Konsequenzen geknüpft. Allerdings beschäftigen sich viele Experimente und Befragungen mit Themen, die entscheidungsabhängige Auszahlungen gar nicht zulassen. Wenn es z.B. um die Zahlungsbereitschaft für das Vermeiden zukünftiger Umweltschäden geht, werden finanzielle Konsequenzen erst dann spürbar, wenn das Programm von der öffentlichen Hand tatsächlich umgesetzt wird und zu dessen Finanzierung neue Steuern erhoben werden. In SP-Befragungen muss daher die Experimentleitung versuchen, die Teilnehmenden davon zu überzeugen, dass deren Entscheidungen tatsächlich einen Einfluss auf die Realisierung des Projekts haben, und die Fragen so stellen, dass kein Anreiz zu strategischem Verhalten besteht. Es ist allerdings schwierig abzuschätzen, wie erfolgreich die Experimentleitung tatsächlich jeweils dabei ist.

Es wurden bislang Tausende von SP-Studien durchgeführt. Man würde erwarten, dass viele davon nachträglich einer Evaluation unterzogen wurden, ob denn die prognostizierten Ergebnisse tatsächlich in der Realität eingetreten sind. Nur ein Vergleich von geäußerten Präferenzen mit tatsächlichem Verhalten ermöglicht eine verlässliche Abschätzung des Hypothetical bias. Nachdem SP-Befragungen vor allem dann durchgeführt werden, wenn es keine oder noch keine Marktdaten gibt, wird oft versucht, die Validität der Befragungsdaten mit anderen Methoden abzuschätzen. Eine Möglichkeit besteht im Test auf Robustheit mittels Replikation der Ergebnisse durch weitere Studien, eine andere im Vergleich mit zusätzlichen Quellen: Beispielsweise werden bei der in der Umweltökonomie häufig verwendeten „Travel cost method" Reisekosten herangezogen, welche die Besucher eines Naherholungsgebiets auf sich nehmen, um die Zahlungsbereitschaft für dieses Naherholungsgebiet zu ermitteln (Whitehead et al. 2008). Manchmal stehen auch Daten aus Experimenten mit monetären Anreizen zum Vergleich zur Verfügung. Von Marketingstudien abgesehen (Morwitz et al. 2007), gibt es wenige Untersuchungen, die auf Marktdaten zurückgreifen können. Wenn es um Existenzwerte geht, ist dies auch schlechthin unmöglich.

Metastudien zur Kontingenten Bewertung zeigen, dass die Zahlungsbereitschaft in hypothetischen Befragungen oft überschätzt wird (List und Gallet 2001; Little et al. 2004; Murphy et al. 2005), wobei die Überschätzung 300 Prozent und mehr betragen kann. Der Schätzfehler ist tendenziell für private Güter weniger groß als für öffentliche Güter, hängt aber auch von vielen anderen Details der Befragungsmethoden ab (Harrison und Rutström 2008). Rakotonarivo et al. (2016) evaluieren 107 neuere Studien zu Discrete Choice Experimenten für Umweltgüter, die Aussagen zu Verlässlichkeit und Validität in irgendeiner Form beinhalten. Sie finden nur 11 Publikationen, die einen Vergleich von hypothetischen und realen Entscheidungen durchführen. Diese identifizieren einen Hypothetical bias im Ausmaß von 50 bis 100 Prozent. Beispiele aus der Verkehrsökonomie zeigen in die andere Richtung: Die Zahlungsbereitschaft zur Einsparung von Reisezeit ist real höher, als hypothetische geäußerte Präferenzen erwarten

lassen (Brownstone und Small 2005). Es gibt jedoch auch eine Zahl von Studien, die keine Evidenz für Hypothetical bias finden (Vossler und Evans 2009; Vossler et al. 2012).

Eine mögliche Ursache des Hypothetical bias wird darin gesehen, dass die Befragungsteilnehmenden über etwas befinden müssen, mit dem sie wenig vertraut sind. Allerdings treten signifikante Unterschiede zwischen geäußerten Präferenzen und tatsächlichem Verhalten nicht nur bei Gütern auf, mit denen die Befragten relativ wenig vertraut sind (beispielsweise gentechnisch veränderte Lebensmittel (Dannenberg et al. 2009)), sondern auch bei Gütern des alltäglichen Gebrauchs (z.B. Beefsteaks (Lusk und Schroeder 2004), Kabelfernsehen (Hausman 2012)).

Ein besonders eindrückliches Beispiel eines Hypothetical bias einer Conjoint Analyse wird in den Abbildungen 3 und 4 gezeigt (Sichtmann et al. 2011). Die Befragten mussten in einer hypothetischen Situation die Zahlungsbereitschaft zwischen verschiedenen Schokoladensorten angeben. Die mit „CBCA" bezeichnete Kurve in Abbildung 3 zeigt die Anzahl der Personen, die angegeben haben, „Hachez Milchschokolade" zu einem bestimmten Preis kaufen zu wollen. Als dieselben Personen dieses Produkt tatsächlich kaufen sollten, waren wesentlich weniger Personen bereit, die Schokolade zu den jeweils vorgegebenen Preisen zu kaufen („BDM"-Kurve). Dieser Unterschied ist nicht nur insgesamt feststellbar, sondern auch auf der Ebene der individuellen Zahlungsbereitschaften, die nur eine sehr schwache Korrelation zwischen geäußerter und tatsächlicher Zahlungsbereitschaft aufweisen (Abb. 4).

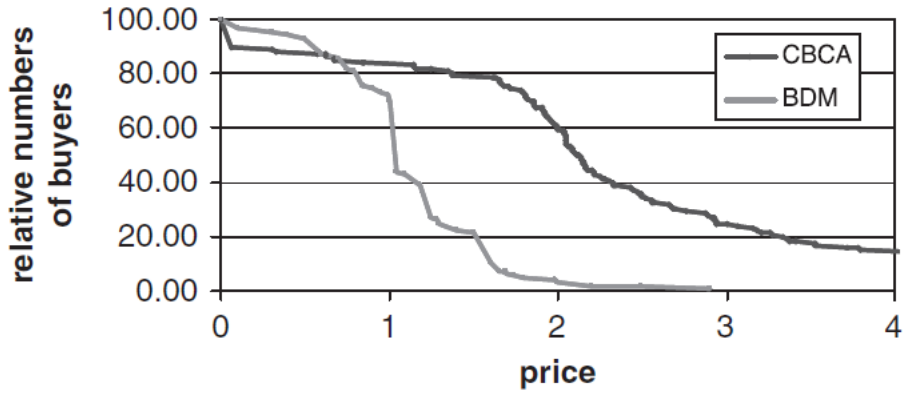

Abbildung 3: Die Anzahl der hypothetischen Käufer (CBCA) für Schokolade liegt deutlich über den Werten der tatsächlichen Käufer (BDM) (Sichtmann et al. 2011, S. 638)

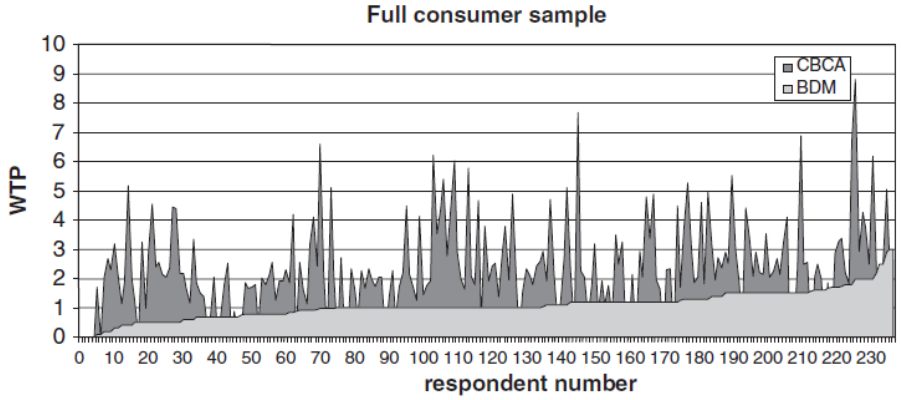

Abbildung 4: Innerhalb des gleichen Individuums variiert der Zusammenhang der hypothetischen (CBCA) und tatsächlichen (BDM) Zahlungsbereitschaft (WTP) stark (Sichtmann et al. 2011, S. 636)

Der Hypothetical bias bezieht sich also auf den Umstand, dass Befragte ihre Entscheidungen in hypothetischen Situationen fällen und keine realen Konsequenzen für ihre Entscheidungen tragen. Es ist daher zweifelhaft, ob Anreizverträglichkeit immer gegeben ist. Louviere listet aber noch andere Themenbereiche auf, die großen Einfluss auf die Realitätsnähe von Befragungen haben: Das Format der Fragestellung und der Kontext, auf den sich die Fragen beziehen.

• Format der Fragestellung und Kontext

Die Aufnahme dieser Punkte in den Katalog beruht auf einer fundamentalen Einsicht, wie sie Hensher (2015) auf Seite 19 formuliert: „The most complex element [...] is the human being". Menschliche Entscheidungen sind das Resultat komplexer Prozesse, in denen viele Faktoren eine Rolle spielen. In seiner Zusammenschau der Ergebnisse der Entscheidungsforschung aus vielen Jahrzehnten unterscheidet Nobelpreisträger Kahneman (2012) zwei Typen von Entscheidungsprozessen. Das sogenannte „System 1" trifft schnelle intuitive Entscheidungen, das „System 2" agiert langsamer und überlegt. System-1-Entscheidungen werden ohne große Anstrengungen getroffen, benutzen nur die unmittelbar vorhandenen Informationen und beruhen oft auf Intuition, Gewohnheiten und Heuristiken. Diese Form der Entscheidungen kann daher auch fehleranfällig sein. Viele Entscheidungen im Alltag sind dem System 1 zuzurechnen, wie z.B. die Wahl der Wegstrecke zum Arbeitsplatz.

Für unsere Überlegungen zur Realitätsnähe von SP-Befragungen ist besonders wichtig, dass viele Menschen nur jene Informationen berücksichtigen, die ihnen unmittelbar zur Verfügung stehen, und selten aktiv nach zusätzlichen Informationen suchen. Bildlich gesprochen bekommt das, was „auf dem Tisch" liegt, große Aufmerksamkeit, während nicht präsente, aber vielleicht wesentlich wichtigere Faktoren die Entscheidung gar nicht beeinflussen. Betrachten wir eine typische Entscheidungssituation in einem

Discrete Choice Experiment zur Verkehrsmittelwahl, wie sie in Abbildung 2 dargestellt ist.

Versetzen Sie sich selbst in die Lage, zwischen einer Fahrt mit dem Auto oder mit dem öffentlichen Verkehr wählen zu müssen. Als Erstes stellt sich doch die Frage, wann Sie denn überhaupt vor einer solchen Wahl stehen könnten und nicht einfach aus Gewohnheit ein bestimmtes Transportmittel nehmen würden (Verplanken et al. 1997). Die dargestellte Entscheidungssituation ist bar jeglichen Kontextes: Sie wissen nicht, zu welchem Zweck (Einkauf, Arbeitsplatz, Freizeit) und zu welcher Tageszeit Sie die Fahrt unternehmen. Sie werden auch nicht darüber informiert, ob Sie direkt an den Ausgangspunkt zurückkehren oder mehrere Ziele anfahren müssen. In der Praxis wird manchmal versucht, diesem Problem durch einmalige Breitstellung relevanter Informationen zu Beginn der Befragung zu begegnen. Aber sind diese Informationen noch präsent, wenn Sie unzählige Entscheidungsblätter zu bearbeiten haben, bei denen sich Ihre Aufmerksamkeit auf die unmittelbar dargestellte Information richten sollte? Andererseits werden Ihnen in den Entscheidungsblättern Informationen vorgelegt, die Sie wahrscheinlich in einer realen Situation nur sehr beschränkt oder gar nicht zur Verfügung haben. Kennen Sie Ihre Benzinkosten, wenn Sie für den Wocheneinkauf zum Supermarkt fahren? Kennen Sie Ticketpreise, Tramlinien, Fahr- und Wartezeiten, wenn Sie nicht regelmäßig dieselbe Strecke mit dem öffentlichen Verkehr zurücklegen? Die Befragungsteilnehmenden äußern also ihre Präferenzen in einer Situation, die sie so in der Realität nicht antreffen werden. Sie werden wahrscheinlich auf Zeit- und monetäre Kosten überreagieren, weil das die einzigen Informationen „auf dem Tisch" sind. Wir präsentieren in Abschnitt 3 experimentelle Evidenz zur geringen Aussagekraft von SP-Befragungen, wenn relevante Kontexte fehlen.

Unserer Meinung nach besteht ein Hauptproblem hypothetischer Fragen darin, dass es psychologisch ungemein schwierig ist, sich mental in eine reale Entscheidungssituation hineinzuversetzen und die eigene Entscheidung zu prognostizieren, wenn die Entscheidung keine Konsequenzen nach sich zieht. Diese mentale Leistung muss im Wesentlichen vom System 2 erbracht werden, selbst wenn die tatsächlichen Entscheidungen im realen Umfeld von System 1 gesteuert werden. Der Entscheidungsmodus in der Befragung ist also mit dem Entscheidungsmodus in der Realität nicht kompatibel. Das Ausfüllen von langen Fragebögen, sei es per Computer oder auf Papier, macht diese Aufgabe nicht gerade einfacher.

Ein ganz allgemeines Problem von SP-Befragungen, und selbst auch von Experimenten mit monetären Anreizen, besteht darin, dass meist eine große Anzahl an Wahlhandlungen zur Schätzung von Modellparametern benötigt wird. Daher werden viele ähnliche Entscheidungssituationen präsentiert, die sich nur in wenigen Merkmalsausprägungen unterscheiden. Die vielfachen Wiederholungen verlangen von den Befragten ein hohes Maß an Konzentration. Dies kann zu Ermüdungserscheinungen sowie Motivationsverlust führen. Wie Unterschiede in den

Merkmalsausprägungen wahrgenommen werden, ist oft nicht nachprüfbar (Ariely und Norton 2008; Carlsson 2010). Darüber hinaus kann auch ein Konflikt zwischen der Komplexität einer Entscheidungssituation und den kognitiven Anforderungen bestehen, wenn die Produkte durch viele verschiedene Merkmale charakterisiert werden. Aus diesem Grund werden oft nur wenige Merkmale präsentiert, was auf Kosten der Realitätsnähe gehen kann.

Die Präferenzforschung ist sich des Hypothetical bias wohl bewusst und hat eine Reihe von Maßnahmen vorgeschlagen, wie man diesem Problem begegnen könnte.

2.4 Methoden zur Begrenzung des Hypothetical bias

In der Literatur werden verschiedene Ansätze diskutiert, wie Fehleinschätzungen aufgrund von Hypothetical bias verringert werden können. Diese können in ex ante Ansätze zur Gestaltung der Befragung und in ex post Ansätze zur Datenauswertung unterteilt werden (Loomis 2014). Ex ante können Entscheidungssituationen realitätsnäher gestaltet werden, wenn mögliche Konsequenzen der Entscheidungen für die Befragten, wie z.B. die zusätzliche Steuerbelastung zur Finanzierung eines Umweltprojekts, deutlich herausgestrichen werden. Außerdem kann ein direkter Appell, so ehrlich und realistisch wie möglich zu antworten, den Bias eventuell verringern. In der Befragung von Cummings und Taylor (1999) erwies sich sogenannter „Cheap talk" als effektiv: Teilnehmende wurden darauf hingewiesen, dass Befragte in früheren Umfragen tendenziell eine zu hohe Zahlungsbereitschaft angegeben haben. Sie wurden weiters darum gebeten, dies zu vermeiden und so wahrheitsgetreu wie möglich zu antworten. Man kann auch versuchen, die Befragten zur Wahrheit zu verpflichten, indem man sie vor der Befragung zu einer eidesstattlichen Erklärung auffordert (Blommaert et al. 2009).

Eine wichtige Kategorie möglicher Verzerrungen stellt der „Social desirability bias" (soziale Erwünschtheit) dar. Diese Verzerrung steht für die Tendenz, so zu antworten, wie es die Gesellschaft oder der Experimentleiter mutmaßlich erwarten. Eine Möglichkeit, diese Art des Hypothetical bias zu verringern, besteht darin, die Teilnehmenden nicht nach dem eigenen Verhalten zu befragen, sondern nach ihrer Meinung zum Verhalten der Mitmenschen (Lusk und Norwood 2009).[2]

Prelec (2004) entwickelte eine vielversprechende Methode, um die Befragten zu wahrheitsgemäßen Antworten zu veranlassen. Sein „Bayesianisches Wahrheitsserum" beruht auf der Idee, dass die Befragten ihre eigene Meinung als Evidenz für die Verteilung der Meinungen in der Bevölkerung interpretieren. So wird wahrscheinlich eine Person, die selbst Steuern hinterzieht, davon ausgehen, dass der Anteil der Steuerhinterzieher vergleichsweise hoch ist. Das Wahrheitsserum macht sich diese Beziehung zunutze, indem „überraschend häufige" Antworten

[2] Das Problem der sozialen Erwünschtheit ist besonders virulent, wenn heikle Themen, wie z. B. Drogenkonsum oder bestimmte Sexualpraktiken, Ziel der Untersuchung sind. Um in diesen Fällen Anreize zu wahren Antworten zu schaffen, werden verschiedene Methoden wie die „Randomized Response Techniques" oder „List Randomization" eingesetzt. Aber auch hier gibt es kontroverse Meinungen, ob diese Verfahren tatsächlich besser als direkte Fragen nach dem Verhalten sind (John et al., 2016; Lensvelt-Mulders et al. 2005).

belohnt und „überraschend seltene" Antworten bestraft werden. Das Verhältnis der tatsächlichen relativen Häufigkeit einer bestimmten Antwort zum (geometrischen) Mittel der von den Befragten geschätzten Häufigkeiten ergibt eine Kennzahl, die als Basis der finanziellen Entlohnung der Befragten dient. Weaver und Prelec (2013) zeigen, dass mit diesem Verfahren der Hypothetical bias einer Kontingenten Bewertung eliminiert werden kann. In einer realen Abstimmung über eine Spende an eine wohltätige Organisation stimmten 44 Prozent der Befragten dafür, in der hypothetischen Situation waren es 76 Prozent, in der hypothetischen Situation mit Wahrheitsserum wiederum nur 47 Prozent. In ihrer Studie finden Barrage und Lee (2010) allerdings nur in einem der beiden von ihnen getesteten Szenarien positive Effekte. Die Herausforderung bei dieser Methode besteht darin, die Studienteilnehmenden von der Wirksamkeit des Mechanismus zu überzeugen, ohne auf die technischen Details einzugehen, die für Laien schwer verständlich sind.

Ein weiterer Vorschlag, die Realitätsnähe hypothetischer Entscheidungssituationen zu verbessern, besteht darin, die Ausprägungen der Produktmerkmale nicht zu weit von bisher tatsächlich erfahrenen Werten abweichen zu lassen (Hensher 2010). In der Mobilitätsforschung ist es herrschende Praxis, vorgängig zu einer SP-Befragung Daten zum tatsächlichen Mobilitätsverhalten zu erheben. Dabei werden die Befragten gebeten, Angaben zu Häufigkeit, Zeitdauer und Verkehrsmittelwahl für bestimmte Fahrten zu machen. Die so gewonnenen Daten werden dann zur Erstellung individualisierter Entscheidungsblätter verwendet. Nun beruhen diese Daten wiederum nicht auf tatsächlich beobachtetem Verhalten, sondern nur auf den Angaben der Befragten, die mehr oder weniger wahrheitsgetreu sein können. Das Problem des Hypothetical bias wird also damit nicht gelöst. Entgegen der gängigen Praxis sollten unserer Meinung nach so gewonnene Daten nicht als „offenbarte Präferenzen" (Revealed preferences) bezeichnet werden. Dieser Begriff ist in den Wirtschaftswissenschaften für tatsächlich objektiv beobachtbares Verhalten reserviert.

Auch ex post lässt sich unter Umständen der Hypothetical bias und eine damit einhergehende Verzerrung der Analyseergebnisse verringern. Durch Screening der Daten mit Hilfe statistischer Methoden kann der Einfluss von Ausreißern verringert und somit ein ausgewogeneres Ergebnis erzielt werden (Carson und Mitchell 1989). Davies und Loomis (2010) schlagen beispielsweise vor, Zahlungsbereitschaften, die höher als der Median sind, als ordinale Indikatoren statt als kardinale Werte zu interpretieren, und die gemessenen Zahlungsbereitschaften entsprechend anzupassen. Falls sowohl hypothetische als auch reale Entscheidungen vorliegen, kann man Kalibrierungsfunktionen schätzen, indem man die Differenz der gemessenen Werte auf beobachtbare Charakteristika der Befragten regressiert. Die Ergebnisse dieses Verfahrens können allerdings nicht auf andere Güter und Kontexte übertragen werden (List und Shogren 1998; Fox et al. 1998). So ist denn der Vorschlag der NOAA, als Faustregel hypothetische Werte zu halbieren, mit Vorsicht zu genießen. Eine weitere Möglichkeit, verzerrte Angaben herauszufiltern, ist die zusätzlich gestellte Frage, wie sicher sich die Befragten ihrer Antworten sind. Geben die Befragten relativ große Unsicherheit an, so wird die Antwort nicht mit in die Analyse einbezogen (Champ et al. 2009). Allerdings besteht kein Konsens

darüber, wie effektiv all diese Maßnahmen tatsächlich sind und welcher Maßnahme der Vorzug gegeben werden soll (Haab et al. 2013; Rakotonarivo et al. 2016).

Schließlich können Daten zu offenbarten Präferenzen, sogenannte Revealed preference (RP-)Daten, in die Analyse einbezogen werden. Im Gegensatz zu SP-Befragungen, welche auf Verhaltensabsichten und Antworten in hypothetischen Situationen beruhen, basieren Revealed preference Daten auf tatsächlichem Verhalten oder beobachteten Marktentscheidungen (Ben-Akiva et al. 1994). Daten von RP-Befragungen werden daher nicht mittels Online- oder Telefonumfragen, sondern im Feld oder in Laborexperimenten erhoben, in welchen das Verhalten der Teilnehmenden unter Anreizen beobachtet werden kann.

Es gibt allerdings auch Nachteile von RP-Befragungen. Zum einen ist die Stichprobe der Teilnehmenden einer RP-Befragung auf aktuelle KonsumentInnen beschränkt (Whitehead et al. 2008). Z.B. kann bei einer Untersuchung zur Verwendung von öffentlichen Verkehrsmitteln nur das Verhalten von denjenigen Menschen erfasst werden, die auch tatsächlich öffentliche Verkehrsmittel verwenden, und nicht von denjenigen, die sie unter gewissen Umständen verwenden würden. Es gibt aber auch noch einen weiteren, entscheidenden Nachteil von RP-Befragungen. RP-Befragungen können sich nur auf Veränderungen beziehen, welche bereits erfolgt sind (Bradley und Kroes 1992). So können RP-Befragungen daher z.B. nur die ex-post-Zahlungsbereitschaft für bestehende Güter und Dienstleistungen schätzen, während SP-Befragungen die ex ante Zahlungsbereitschaft für neue Güter und Dienstleistungen abfragen können (Whitehead et al. 2008). Dies ist im Kontext politischer Maßnahmen oder wirtschaftlicher Entscheidungen relevant. Der Möglichkeit, ausschließlich RP-Daten zu verwenden, steht in der Praxis entgegen, dass diese häufig zu wenig Variabilität und zu hohe Korrelationen zwischen den Ausprägungen der Produktmerkmale aufweisen. Diese Schwächen sind oft mit ein Grund, warum man unter Umständen auf SP-Befragungen zurückgreifen muss (Kroes und Sheldon 1988; Vrtic 2005). Als Lösung dieser Probleme wird daher die Kombination von SP-Daten mit RP-Daten vorgeschlagen, sofern solche vorhanden sind (Ben-Akiva et al. 1994). Diese Vorgehensweise wird heute von vielen Forschern als State-of-the-Art-Methode angesehen (Whitehead et al. 2008; Ben-Akiva et al. 1994).[3]

3 Evidenz zur Kontextabhängigkeit von Entscheidungen

Wie aus dem Zitat von Louviere hervorgeht (Louviere 2006), ist die Realitätsnähe einer Entscheidungssituation unter anderem maßgeblich davon abhängig, in welchem Kontext die Entscheidung getroffen wird. Im Folgenden diskutieren wir zwei neuere Untersuchungen bezüglich privater und öffentlicher Güter, deren Ergebnisse ganz klar die Kontextabhängigkeit von SP-Befragungen zeigen. Die erste Studie von Fehr et al. (2018) beschäftigt sich mit der Verkehrsmittelwahl im Einkaufsverkehr in

[3] Auch hier gilt unsere Warnung von vorhin: Angaben, die nicht auf tatsächlich beobachtetem Verhalten sondern nur auf Befragungen der Studienteilnehmer beruhen, sollten nicht als RP-Daten klassifiziert werden.

der Schweiz, die zweite Studie von Tinch et al. (2015) untersucht die Zahlungsbereitschaft für Landschaftsmaßnahmen in einem britischen Nationalpark.

3.1 Choice Experiment zur Verkehrsmittelwahl

Das Experiment von Fehr et al. (2018) wurde mit einer für die deutschsprachige Schweiz repräsentativen Stichprobe von 1.504 Personen online durchgeführt. Voraussetzung für die Teilnahme war die Verfügbarkeit eines PKW und der Besitz eines Führerscheins. Nach einer Einführung durchliefen alle Teilnehmenden fünf verschiedene Entscheidungssituationen, welche sie jeweils zur Wahl des Verkehrsmittels (PKW oder öffentliches Verkehrsmittel – ÖV) zur Erreichung ihres Einkaufsortes befragten. Die monetären Kosten und die benötigte Zeit (im Folgenden summarisch „Kosten" genannt) für beide Verkehrsmittel wurden dabei jeweils gegenübergestellt. Ein beispielhaftes Entscheidungsblatt ist in Abbildung 5 dargestellt.

Abbildung 5: Beispiel Entscheidungssituation: Exemplarische Entscheidungssituation "ohne Kontext" (Fehr et al. 2018)

In dieser Studie ging es vor allem um die Untersuchung des Einflusses von Kontextfaktoren auf die Kostensensitivität, und nicht um die Messung von Trade-offs zwischen den einzelnen Komponenten. Zu diesem Zweck wurden Zeit- und Kostenfaktoren folgendermaßen variiert: Bei Kostenstufe 1 war man mit dem Auto bedeutend schneller am Ziel und die monetären Kosten für die Autofahrt waren sehr viel geringer als die Kosten für den ÖV. In der zweiten Kostenstufe waren noch immer leichte Vorteile des Autos dargestellt. Die dritte Kostenstufe zeichnete sich dadurch aus, dass Auto und ÖV in Bezug auf Zeit- und monetäre Attribute vergleichbar waren. In der vierten Kostenstufe war der ÖV insgesamt leicht, in der fünften wesentlich günstiger als das Auto.

Während die Kontrollgruppe ausschließlich über diese Kosteninformationen verfügte, erhielt die Versuchsgruppe zusätzlich Angaben zum Kontext, in welchem sie ihren Einkauf durchführte. So konnte das Wetter gut oder schlecht sein, ein kleiner oder großer Einkauf anstehen, und die Fahrt nach dem Einkaufen entweder zu einem weiteren Ort fortgesetzt werden oder direkt zurück zum Ausgangspunkt erfolgen. Die Teilnehmenden erhielten jeweils eine zufällig zugeteilte Kontextbeschreibung für jede der Entscheidungssituationen, sodass über alle Teilnehmenden hinweg jede Entscheidungssituation mit jeder Kontextbeschreibung kombiniert werden konnte.

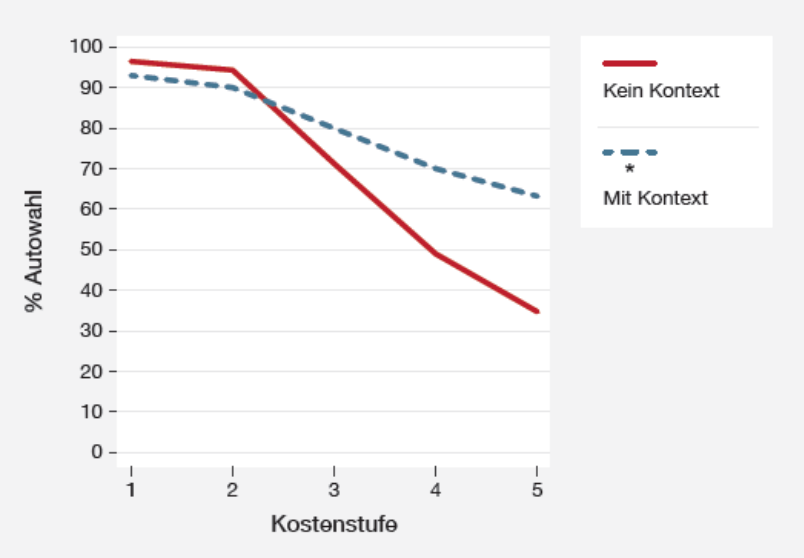

Abbildung 6: Sensitivität der Verkehrsmittelwahl bei Kostenveränderung in Situationen mit und ohne Kontextvariablen (Fehr et al. 2018)

Abbildung 6 demonstriert das Hauptresultat der Studie. Die Grafik zeigt den Prozentsatz der gewählten Autofahrten in Abhängigkeit von der Kostenstufe. Die durchgezogene Linie bezieht sich auf die Entscheidungen der Kontrollgruppe (ohne Kontext), die gestrichelte auf die Entscheidungen der Versuchsgruppe. Wenn die Fahrt mit dem Auto günstiger ist (Kostenstufen 1 und 2), wird in mindestens 90 Prozent der Entscheidungen das Auto gewählt, und zwar unabhängig vom Kontext der Entscheidung. Verursachen jedoch Auto und ÖV in etwa dieselben Gesamtkosten, sollten die Befragten im Wesentlichen indifferent sein, wenn Zeit und monetäre Kosten die ausschließlichen Treiber des Verhaltens wären. Es zeigt sich aber, dass ungefähr 70 Prozent der Entscheidungen der Kontrollgruppe zugunsten des Autos ausfallen. In der Versuchsgruppe sind dies sogar ca. 80 Prozent. Ist der ÖV kostengünstiger, nimmt diese Diskrepanz noch deutlich zu – kontextbezogene Entscheidungen fallen wesentlich häufiger zugunsten des Autos aus, als abstrakte, kontextlose Entscheidungen. Das bloße Vorhandensein eines Entscheidungskontextes führt also zu einer massiv geringeren Sensitivität in Bezug auf die Kosten der Verkehrsmittel. Anders ausgedrückt: Die abstrakte Darstellung der Entscheidungssituation, bei der nur monetäre Kosten und Zeitaufwand für die Befragten unmittelbar verfügbar sind, hat eine vergleichsweise Überreaktion auf die Kostendifferenzen zur Folge. Dieses einfache Experiment zeigt, wie wenig belastbar

Ergebnisse von SP-Befragungen selbst für alltägliche private Güter sind, wenn die Befragten mit abstrakten und kontextunabhängigen, und daher wenig realitätsnahen, Entscheidungssituationen konfrontiert werden.

3.2 SP-Befragung zu Landschaftsmaßnahmen in einem Nationalpark

Tinch et al. (2015) führten ein Discrete Choice Experiment durch, um die Zahlungsbereitschaft für Maßnahmen zur Änderung der Nutzungsintensität, und damit einhergehend der Biodiversität, von verschiedenen Landschaften im Peak District National Park in Großbritannien zu erheben. Das Experimentdesign umfasste vier verschiedene Versuchsbedingungen mit denselben 52 Befragungsteilnehmenden, wobei die vorgelegten Szenarien immer identisch waren. Die vier Versuchsbedingungen unterschieden sich in Ort und Zeitpunkt der Befragung:

1. In Versuchsbedingung 1 fand die Befragung vor Besuch des Nationalparks an einem Ort außerhalb des Parks statt.

2. In Versuchsbedingung 2 wurden die Teilnehmenden in den Park gebracht und konnten sich vor Ort ein Bild machen. Sie wurden auch direkt im Park befragt.

3. In Versuchsbedingung 3 erfolgte die Befragung unmittelbar nach Rückkehr von der Besichtigungstour.

4. Die 4. Befragung fand 4 Monate nach Besichtigung an einem Ort außerhalb des Parks statt.

Ganz allgemein zeigten die Befragten eine Abneigung gegen Änderungen des Status quo, d.h. dass Veränderungen in Richtung mehr oder weniger intensivem Landschaftsmanagement mit negativen Zahlungsbereitschaften verbunden waren. Am stärksten negativ reagierten die geschätzten Zahlungsbereitschaften in Versuchsbedingung 1 auf Änderungen der Nutzungsintensität, also bevor die Teilnehmenden dem Park einen Besuch abgestattet hatten. Die verblüffendsten Ergebnisse betreffen aber Versuchsbedingung 2, in der die Befragten direkt vor Ort ihre Entscheidungen treffen mussten: Wenn sie die verschiedenen Landschaften inspizieren konnten, fiel es den Befragten viel schwerer, sich zwischen den Szenarien zu entscheiden. Darüber hinaus verlor die Kostenvariable (die mit den Maßnahmen jeweils einhergehende Steuerbelastung) jegliche statistische Signifikanz. Das bedeutet, dass den Kosten keinerlei Aufmerksamkeit geschenkt wurde und eine Schätzung der Zahlungsbereitschaft unmöglich war. In diesem Experiment hatte der Kontext der Entscheidungssituation offenbar einen dramatischen Effekt auf die geäußerten Präferenzen. Die Studienautoren interpretieren dieses Resultat dahingehend, dass Entscheidungen hauptsächlich von jenen Variablen abhängen, welche die Befragten direkt beobachten können. Allerdings unterscheiden sich auch die Schätzwerte für Versuchsbedingung 1 substanziell von jenen für Versuchsbedingung 3 und 4, die weniger negativ ausfielen.

Das heißt, dass die Abneigung gegen Änderungen des Status quo anfangs am stärksten war und die Erfahrungen im Nationalpark auch die in zeitlicher und räumlicher Distanz getroffenen Entscheidungen beeinflussten.

Dieses Experiment zeigt, dass geäußerte Präferenzen zum Zeitpunkt des Konsums eines öffentlichen Gutes massiv von jenen in einer abstrakten Entscheidungssituation abweichen können. Im Falle von passiven Nutzwerten wird das betreffende Umweltgut gar nie konsumiert, sodass man sich auf herkömmliche SP-Befragungen stützen muss. Hat es in diesem Bereich Fortschritte gegeben?

4 Ein Modellbeispiel der Kontingenten Bewertung?

Die durch den Unfall der Exxon Valdez verursachte Ölpest ist nicht die einzige und leider auch nicht die größte geblieben. 2010 geriet die Bohrplattform Deepwater Horizon von BP im Golf von Mexiko in Brand und ging unter. Die dadurch verursachte Ölpest war etwa 20-mal größer als jene des Exxon Valdez Unfalls (Kling et al. 2012). Bishop et al. (2017) berichten, dass die gerichtliche Einigung Zahlungen in Höhe von 20,8 Milliarden Dollar vorsieht, welche die Verluste von direktem Nutzwert und Existenzwert abdecken sollen. Wie könnte man die Höhe dieser Zahlung rechtfertigen?

Bishop et al. (2017) führten eine für die USA repräsentative SP-Befragung durch, die auf die Messung der Zahlungsbereitschaft für die Vermeidung zukünftiger Katastrophen wie jene durch den Untergang der Deepwater Horizon hervorgerufenen abzielt.[4] Die Teilnehmenden mussten über die Realisierung eines bestimmten Programms abstimmen und wurden über Folgendes informiert:

1. den Zustand des Golfs von Mexiko vor dem Unfall,

2. die Ursachen des Unfalls,

3. die eingetretenen Umweltschäden,

4. detaillierte Beschreibung eines Programms zur Vermeidung eines ähnlichen Unfalls, sowie

5. die Kosten des Programms in Form höherer Steuern, falls das Programm verwirklicht würde.

Verschiedenen Gruppen von Teilnehmenden wurden jeweils per Zufall unterschiedliche Projektgrößen und Steuerkosten zugeordnet. Es wurde großer Wert daraufgelegt, die Teilnehmenden von der Wichtigkeit ihrer Entscheidung für die Politikgestaltung und der finanziellen Konsequenzen für sie selbst zu überzeugen. So wurde vor der Durchführung der Befragung ein entsprechender Brief des U.S. Department of Commerce versandt, zu dem die NOAA gehört, welche die

[4] Die Ergebnisse dieser Studie wurden 2017 in der renommierten Wissenschaftszeitschrift Science publiziert.

Federführung im Verfahren gegen BP übernommen hatte. Außerdem wurde der Inhalt des Briefes vor den persönlichen Interviews nochmals besprochen. In der Befragung wurde betont, dass die Umsetzung des Programms zusätzliche Steuereinnahmen erfordere.

Die Studienautoren argumentieren, dass hypothetische Entscheidungen den Kriterien der ökonomischen Rationalität genügen können: Erstens zeigen ihre Ergebnisse, dass bei gegebener Steuerbelastung die Zustimmung mit dem Schadensausmaß zunimmt. Zweitens nimmt bei gegebener Schadenshöhe die Zustimmung ab, wenn die Steuerbelastung erhöht wird. Beide Effekte nehmen plausible Größenordnungen an, sodass die Studienautoren dem Hausmanschen Hauptvorwurf der mangelnden Mengensensitivität begegnen können. Bishop et al. (2017) schätzen auf Basis der Antworten die untere Grenze der Zahlungsbereitschaft für die zukünftige Vermeidung einer Ölpest auf 17,2 Milliarden US Dollar. Dieser Wert floss unter anderem auch in die letztendliche außergerichtliche Einigung über die Schadenshöhe ein.

Die beiden erwähnten Rationalitätskriterien stellen allerdings nur eine notwendige Bedingung für die Glaubwürdigkeit der Resultate dar. Damit ist noch nicht garantiert, dass kein Hypothetical bias vorliegt.

5 Schlussfolgerungen: Größere Validierungsanstrengungen erforderlich

Die teilweise heftige Kritik an der Kontingenten Bewertungsmethode und anderer SP-Befragungsformate wegen ihrer mangelnden Glaubwürdigkeit und Verlässlichkeit hat in vielen Bereichen zu einem Umdenken geführt. In seinem Hauptvortrag an der 10. internationalen Konferenz zu Befragungsmethoden in der Verkehrsökonomie 2014 betonte Hensher (2015) die Bedeutung verhaltenswissenschaftlicher Einsichten für die Gestaltung von realitätsnahen Befragungen. Trotzdem sehen wir noch großen Handlungsbedarf. Unseres Erachtens gibt es keinen verlässlichen Standard, anhand dessen man Richtung und Ausmaß des Hypothetical bias in einer konkreten Befragung ex ante abschätzen könnte. Es werden aber immer noch zu wenige Anstrengungen unternommen, hypothetische Entscheidungen anhand realen Verhaltens zu validieren. Solch ein Unterfangen kostet Geld und bedarf sorgfältiger Vorbereitung und Durchführung mit Kontroll- und Versuchsgruppen. Wann immer möglich, empfehlen wir, zusätzlich zu SP-Befragungen begleitende Labor- oder Feldexperimente durchzuführen, die mit realen monetären Anreizen arbeiten. Volle Realitätsnähe kann man mit Feldexperimenten erreichen, in denen die Teilnehmenden gar nicht wissen, dass sie an einem Experiment teilnehmen.

Da Existenzwerte nicht handelbar sind und daher keinerlei Spuren in Märkten hinterlassen, kann die Größenordnung von Existenzwerten allerdings nur mittels hypothetischer Fragen abgeschätzt werden. In diesen Fällen müssen große Anstrengungen unternommen werden, möglichst anreizverträgliche Entscheidungssituationen zu generieren. Unseres Erachtens sind die Studienergebnisse von Bishop et al. (2017) in dieser Hinsicht ermutigend. Wir schlagen aber auch hier vor, es nicht bei der Betonung von finanziellen

Konsequenzen der Entscheidungen in den Instruktionen zu belassen. Man könnte beispielsweise eine kleine Stichprobe auswählen und dann eine Kontrollgruppe ohne Bayesianisches Wahrheitsserum und eine Versuchsgruppe mit Bayesianischem Wahrheitsserum befragen. Sollte sich eine große Diskrepanz in den Antworten ergeben, könnte man die Ergebnisse der großen Stichprobe, die nur eine SP-Befragung durchmacht, entsprechend kalibrieren.

Auftraggeber, ob private oder öffentliche, sollten sich nicht mit dem Hinweis, dass die vorgeschlagene Studie Best Practice wäre und dem aktuellen Stand der Forschung entspräche, abspeisen lassen, sondern darauf drängen, dass die Auftragnehmer den Nachweis der externen Validität erbringen müssen. Experimente, und insbesondere Feldexperimente, sind im Vergleich zu SP-Befragungen aufwändiger und teurer. Mögliche Folgekosten durch ineffiziente Allokation von Gütern, die durch eine verzerrte Schätzung in SP-Befragungen entstehen, können diese Experimentalkosten jedoch um ein Vielfaches übersteigen. Gerade, wenn es um wirtschaftspolitische Maßnahmen geht, die viele Menschen betreffen, sollten Forscher stets die beste zur Verfügung stehende Methode anwenden. SP-Befragungen ohne begleitende Experimente mit realen Anreizen genügen selten dieser Anforderung. Nur wenn das Ausmaß des Hypothetical bias verlässlich abgeschätzt werden kann, ist eine Zahl besser als keine Zahl.

6 Literaturverzeichnis

Ariely, D. und M.I. Norton (2008), How actions create–not just reveal–preferences. *Trends in Cognitive Sciences* 12(1), S. 13–16.

Barrage, L. und M.S. Lee (2010), A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics letters* 106(2), S. 140–42.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, und V. Rao (1994), Combining revealed und stated preferences data. *Marketing Letters* 5(4), S. 335–49.

Bishop, R.C., K.J. Boyle, R.T. Carson, D. Chapman, W.M. Hanemann, B. Kanninen, R.J. Kopp, J.A. Krosnick, J. List, N. Meade et al. (2017), Putting a value on injuries to natural assets: The bp oil spill. *Science* 356(6335), S. 253–54.

Blommaert, J., M. Baynham, A. De Fina, D. Eades, M. Jacquemet, A. Jaffe, K. Maryns, T. McNamara, R. Moore, S.S. Mufwene et al. (2009), Language, asylum, und the national order. *Current Anthropology* 50(4), S. 415–41.

Botzen, W.W. und J.C. van den Bergh (2012), Risk attitudes to low-probability climate change risks: Wtp for flood insurance. *Journal of Economic Behavior & Organization* 82(1), S. 151–66.

Bradley, M.A. und E.P. Kroes (1992), Forecasting issues in stated preference survey research. *In Selected Readings in Transport Survey Methodology*. Edited Proceedings of the 3rd International Conference on Survey Methods in Transportation, January 5-7 1990, Washington, D.C.

Brownstone, D. und K.A. Small (2005), Valuing time und reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy und Practice* 39(4), S. 279–93.

Carlsson, F. (2010), Design of stated preference surveys: Is there more to learn from behavioral economics? *Environmental und Resource Economics* 46(2), S. 167–77.

Carson, R.T. und T. Groves (2007), Incentive und informational properties of preference questions. *Environmental und resource economics* 37(1), S. 181–210.

Carson, R.T. und R.C. Mitchell (1989), Using surveys to value public goods: the contingent valuation method. *Resources for the Future*, Washington DC, 82.

Carson, R.T., R.C. Mitchell, M. Hanemann, R.J. Kopp, R. S. Presser und P.A. Ruud (2003), Contingent valuation und lost passive use: damages from the exxon valdez oil spill. *Environmental und resource economics* 25(3), S. 257–86.

Champ, P.A., R. Moore und R.C. Bishop (2009), A comparison of approaches to mitigate hypothetical bias. *Agricultural und Resource Economics Review* 38(2), S. 166–80.

Cummings, R.G. und L.O. Taylor (1999), Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. *American Economic Review* 89(3), S. 649–65.

Dannenberg, A., S. Scatasta, und B. Sturm (2009), Keine Chance für genetisch veränderte Lebensmittel in Deutschland? Eine experimentelle Analyse von Zahlungsbereitschaften. *Perspektiven der Wirtschaftspolitik* 10(2), S. 214–34.

Davies, S. und J. Loomis (2010), An improved method for calibrating purchase intentions in stated preference demand models. *Journal of Agricultural & Applied Economics* 42(4), S. 679.

Diamond, P.A. und J.A. Hausman (1994), Contingent valuation: is some number better than no number? *Journal of Economic Perspectives* 8(4), S. 45–64.

Fehr, G., L. Geisseler, und M. Jäger (2018), Der Mensch im Verkehr: Ein Homo Oeconomicus? https://www.zukunft-mobilitaet.ch/images/Verhaltensoekonomie/unitegallery_thumbs/Studie-FehrAdvice_Der-Mensch-im-Verkehr--Kein-homo-oeconomicus.pdf

Fox, J.A., J.F. Shogren, D.J. Hayes und J.B. Kliebenstein (1998), Cvm-x: calibrating contingent values with experimental auction markets. *American Journal of Agricultural Economics* 80(3), S. 455–65.

Haab, T.C., M.G. Interis, D.R. Petrolia und J.C. Whitehead (2013), From hopeless to curious? thoughts on hausman's "dubious to hopeless" critique of contingent valuation. *Applied Economic Perspectives und Policy* 35(4), S. 593–612.

Hammack, J. und G.M. Brown (1974), Waterfowl und wetlands: Toward bioeconomic analysis. Resources for the future, Washington D.C.

Harrison, G.W. und E.E. Rutström (2008), Experimental evidence on the existence of hypothetical bias in value elicitation methods. *Handbook of Experimental Economics Results* 1, S. 752–67.

Hausman, J. (2012), Contingent valuation: from dubious to hopeless. *Journal of Economic Perspectives* 26(4), S. 43–56.

Hensher, D.A. (2010), Hypothetical bias, choice experiments und willingness to pay. *Transportation Research Part B: Methodological* 44(6) S. 735–52.

Hensher, D.A. (2015), Data challenges: more behavioural und (relatively) less statistical–a think piece. *Transportation Research Procedia* 11, S. 19–31.

John, L.K., G. Loewenstein, A. Acquisiti und J. Vosgerau (2016), When und why randomized response techniques (fail to) elicit the truth. Harvard Business School.

Kahneman, D. (2012), *Schnelles Denken, langsames Denken*. Siedler Verlag.

Kling, C.L., D.J. Phaneuf und J. Zhao (2012), From exxon to bp: Has some number become better than no number? *Journal of Economic Perspectives* 26(4), S. 3–26.

Kroes, E.P. und R.J. Sheldon (1988), Stated preference methods: an introduction. *Journal of Transport Economics und Policy,* S. 11–25.

Krutilla, J.V. (1967), Conservation reconsidered. *The American Economic Review* 57(4), S. 777–86.

Lensvelt-Mulders, G.J., J.J. Hox, P.G. Van der Heijden und C.J. Maas (2005), Metaanalysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research* 33(3), S. 319–48.

List, J.A. und C.A. Gallet (2001), What experimental protocol influence disparities between actual und hypothetical stated values? *Environmental und Resource Economics* 20(3), S. 241–54.

List, J.A. und J.F. Shogren (1998), Calibration of the difference between actual und hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization* 37(2), S. 193–205.

Little, J., R. Berrens et al. (2004), Explaining disparities between actual und hypothetical stated values: further investigation using meta-analysis. *Economics Bulletin* 3(6), S. 1–13.

Loomis, J.B. (2014), 2013 WAEA keynote address: Strategies for overcoming hypothetical bias in stated preference surveys. *Journal of Agricultural und Resource Economics,* S. 34–46.

Louviere, J.J. (2006), What you don’t know might hurt you: some unresolved issues in the design und analysis of discrete choice experiments. *Environmental und Resource Economics* 34(1), S. 173–88.

Louviere, J.J., T.N. Flynn und R.T. Carson (2010), Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* 3(3), S. 57–72.

Lusk, J.L. und F.B. Norwood (2009), An inferred valuation method. *Land Economics* 85(3), S. 500–14.

Lusk, J.L. und T.C. Schroeder (2004), Are choice experiments incentive compatible? a test with quality differentiated beef steaks. *American Journal of Agricultural Economics* 86(2), S. 467–82.

Manski, C.F. und D. McFadden (1981), *Structural analysis of discrete data with econometric applications*. MIT Press Cambridge, MA.

Morwitz, V.G., J.H. Steckel und A. Gupta (2007), When do purchase intentions predict sales? *International Journal of Forecasting* 23(3), S. 347–64.

Murphy, J.J., P.G. Allen, T.H. Stevens und D. Weatherhead (2005), A meta-analysis of hypothetical bias in stated preference valuation. *Environmental und Resource Economics* 30(3), S. 313–25.

Pierce, R.M. (1994), Valuing the environment: NOAA’s New Regulations under the Oil Protection Act of 1990. *Pepperdine Land Review* 22(1), S. 167–212.

Portney, P.R. (1994), The contingent valuation debate: why economists should care. *Journal of Economic Perspectives* 8(4), S. 3–17.

Prelec, D. (2004), A bayesian truth serum for subjective data. *Science* 306(5695), S. 462–66.

Rakotonarivo, O.S., M. Schaafsma, und N. Hockley (2016), A systematic review of the reliability und validity of discrete choice experiments in valuing non-market environmental goods. *Journal of Environmental Management* 183, S. 98–109.

Sichtmann, C., R. Wilken und A. Diamantopoulos (2011), Estimating willingness-to-pay with choice-based conjoint analysis–can consumer characteristics explain variations in accuracy? *British Journal of Management* 22(4), S. 628–45.

Tinch, D., S. Colombo und N. Hanley (2015), The impacts of elicitation context on stated preferences for agricultural landscapes. *Journal of Agricultural Economics* 66(1), S. 87–107.

Verplanken, B., H. Aarts und A. Van Knippenberg (1997), Habit, information acquisition, und the process of making travel mode choices. *European Journal of Social Psychology* 27(5), S. 539–60.

Vossler, C.A., M. Doyon und D. Rondeau (2012), Truth in consequentiality: theory und field evidence on discrete choice experiments. *American Economic Journal: Microeconomics* 4(4), S. 145–71.

Vossler, C.A. und M.F. Evans (2009), Bridging the gap between the field und the lab: Environmental goods, policy maker input, und consequentiality. *Journal of Environmental Economics und Management* 58(3), S. 338–45.

Vrtic, M. (2005), Ein hierarchisches („nested “) Logit-Modell für die Analyse kombinierter Stated- und Revealed-Preference-Daten zur Verkehrsmittelwahl. Deutsche Verkehrswissenschaftliche Gesellschaft e. V. (Hrsg.), 12.

Weaver, R. und D. Prelec (2013), Creating truth-telling incentives with the bayesian truth serum. *Journal of Marketing Research* 50(3), S. 289–302.
Weis, C., M. Vrtic, K.W. Axhausen und M. Balać (2016), SP-Befragung 2015 zum Verkehrsverhalten. Retrieved from https://www.are.admin. ch/are/de/home/verkehr-und-infrastruktur/grundlagen-und-daten/ stated-preference-befragung.html.

Weis, C., M. Vrtic, K.W. Axhausen und M. Balać (2016), SP-Befragung 2015 zum Verkehrsverhalten. https://www.are.admin. ch/are/de/home/verkehr-und-infrastruktur/grundlagen-und-daten/ stated-preference-befragung.html.

Whitehead, J.C., S.K. Pattanayak, G.L. Van Houtven, und B.R. Gelso (2008), Combining revealed und stated preference data to estimate the nonmarket value of ecological services: an assessment of the state of the science. *Journal of Economic Surveys* 22(5), S. 872–908.