Die „Sonntagsfrage“, soziale Erwünschtheit und die AfD: Wie alternative Messmethoden der Politikwissenschaft weiterhelfen können

Thomas Gschwend, PhD (Universität Mannheim, Fakultät für Sozialwissenschaften, A5, 6, 68131 Mannheim, gschwend@uni-mannheim.de)
Sebastian Juhl (Universität Mannheim, Sonderforschungsbereich 884, B6, 30-32, 68131 Mannheim, sjuhl@mail.uni-mannheim.de)
Roni Lehrer, PhD (Universität Mannheim, Sonderforschungsbereich 884, B6, 30-32, 68131 Mannheim, lehrer@uni-mannheim.de)

Stichwörter:

Wahlabsicht, Sonntagsfrage, Soziale Erwünschtheit, Vorhersage, AfD

Zusammenfassung:

Die Wahlabsichtsfrage, populärwissenschaftlich auch als „Sonntagsfrage” bezeichnet, wird kritisiert, weil mit ihr der Stimmenanteil der Alternative für Deutschland (AfD) nicht valide zu messen ist. Wir argumentieren, dass alternative Messinstrumente, die Verzerrungen aufgrund von sozialer Erwünschtheit berücksichtigen, besser geeignet sind. Dazu testen wir erstmalig drei alternative Messmethoden – das doppelte Listenexperiment, die Randomisierte-Antwort-Technik und die Weisheit-der-Vielen-Methode – hinsichtlich des geschätzten AfD-Stimmenanteils und vergleichen sie mit der klassischen „Sonntagsfrage“. Unsere Ergebnisse zeigen, dass insbesondere die Weisheit-der-Vielen-Methode eine kostengünstige und gute Erweiterung der politikwissenschaftlichen Fragebatterie ist.

1. Einleitung

> "*Wenn am nächsten Sonntag Bundestagswahl wäre, welche Partei würden Sie dann wählen*?"

Die klassische Frage nach der Wahlabsicht von Befragten, die populärwissenschaftlich auch als „Sonntagsfrage“[1] bezeichnet wird, ist die wohl bekannteste und am häufigsten ausgewertete Frage im Kontext einer Vorwahlstudie. Trotz ihrer enormen wissenschaftlichen und medialen Popularität ist sie aber keineswegs gänzlich unumstritten. In diesem Beitrag geht es uns um das beobachtbare und oft beklagte Phänomen, dass der Anteil der Wählerschaft für die nun erstmals in den Bundestag eingezogene AfD nicht valide mit der „Sonntagsfrage“ zu messen ist (Bergmann u. Diemeier 2017). Wir argumentieren, dass alternative Messinstrumente, die Verzerrungen aufgrund von sozialer Erwünschtheit berücksichtigen (Philips u. Clancy 1972), besser geeignet sind, das AfD-Ergebnis bereits vor der Wahl vorherzusagen. Dazu testen wir drei alternative Messmethoden – das doppelte Listenexperiment, die Randomisierte-Antwort-Technik und die Weisheit-der-Vielen-Methode – hinsichtlich der Akkuratesse ihrer Prognosen und vergleichen sie mit der klassischen „Sonntagsfrage“. Unsere Ergebnisse zeigen, dass insbesondere die Weisheit-der-Vielen-Methode eine kostengünstige und gute Erweiterung der politikwissenschaftlichen Fragebatterie ist.

Die Kritik an der „Sonntagsfrage“ ist nicht neu. Einerseits gilt sie als das Standardinstrument der politischen Meinungsforschung zur Messung der politischen Stimmung (z.B. Groß 2010; Wüst 2003; Roth u. Wüst 2015). Andererseits werden gleichzeitig handwerkliche Schwächen eingestanden, die aus dem hypothetischen Charakter der Frageformulierung resultieren. Darüber hinaus gibt es weitere Defizite, aufgrund derer man davor warnen muss, die Unterstützung einzelner Parteien bei der „Sonntagsfrage“ und die aktuelle politische Stimmung bzw. von ihr abgeleitete Erwartung über Wahlergebnisse gleichzusetzen.

Besonders im Kontext von Wahlen mit starken rechtspopulistischen Parteien ist der Nutzen der „Sonntagsfrage“ für die Prognose des Wahlergebnisses fraglich. Mehrere Studien zeigen, dass der Anteil dieser Parteien nicht akkurat durch klassische Umfragen erfasst werden kann (z.B. Aichholzer et al. 2014; Evans u. Ivaldi 2010; Johann et al. 2016). Dieses Problem führt dazu, dass die Aussagekraft der durch Umfragen gewonnenen Ergebnisse deutlich geschmälert wird. Diese systematischen Schwächen des Messinstruments haben auch Auswirkungen auf den Wahlkampf der Partien, da Bevölkerungsumfragen maßgeblich die strategischen Ausrichtungen von Parteien beeinflussen (z.B. Adams et al. 2004; Kollmann et al. 1992).

Eine Begründung, warum solche Verzerrungen entstehen, liegt darin, dass Befragte in Umfragen die Tendenz zeigen, falsche Angaben zu machen, um sich selbst in einem besseren Licht erscheinen zu lassen, statt wahrheitsgemäß eine sozial unerwünschte Antwort zu geben (Philips u. Clancy 1972). Während es etwa sozial akzeptiert ist wählen zu gehen, und von daher die Angaben in Umfragen zur Wahlbeteiligung die tatsächliche Wahlbeteiligung deutlich übersteigt, sind Angaben zur Wahlabsicht für eine extreme oder populistische Partei in Umfragen zu niedrig, vermutlich weil eine solche Wahlabsicht von vielen Befragten als sozial unerwünscht wahrgenommen wird.

Die Umfrageforschung hat alternative Instrumente entwickelt, um validere Messungen sozial

[1] In Abhängigkeit vom durchführenden Meinungsforschungsinstitut kann die genaue Formulierung der "Sonntagsfrage" leicht abweichen (Groß 2010, S. 48).

unerwünschter Eigenschaften zu ermöglichen. Inwiefern diese alternativen Messtechniken es ermöglichen den Anteil populistischer Parteien trotz des Problems der sozialen Erwünschtheit zu ermitteln, stellt das Erkenntnisinteresse dieses Beitrags dar. Dazu fokussieren wir uns auf den speziellen Fall in einer Umfragesituation die Absicht zu erklären, die AfD bei der Bundestagswahl wählen zu wollen. Vor diesem Hintergrund erläutern wir drei dieser alternativen Methoden konzeptionell, vergleichen sie miteinander und evaluieren, wie gut sie zur Schätzung des Stimmanteils populistischer Parteien vor nationalen Wahlen geeignet sind.

In diesem Beitrag testen wir erstmalig drei Umfragetechniken in einer Studie, inwiefern sie valide Ergebnisse hinsichtlich des tatsächlichen Anteils der AfD-Wähler in der Bevölkerung im Kontext der Wahl zum 19. Deutschen Bundestag liefern können. Wir entwickeln dazu neue Instrumente für eine Bevölkerungsumfrage, basierend auf (1) der Randomisierten-Antwort-Technik (RAT) (Höglinger et al. 2016), (2) einem Listen-Experiment (LE) (Johann et al. 2016) und (3) der Weisheit-der-vielen-Methode (Murr 2017). Wir implementieren fortgeschrittene Varianten dieser Instrumente zwei Monate vor der deutschen Bundestagswahl 2017 in der Juli-Welle des German Internet Panels (GIP) (Blom et al. 2015) und vergleichen sie mit dem Ergebnis der „Sonntagsfrage“ zu diesem Zeitpunkt, mit dem amtlichen Wahlergebnis sowie mit den Vorhersagen anderer Umfragen während der Feldzeit. Allen drei Umfragetechniken ist gemein, dass sie keine Rückschlüsse auf das individuelle Verhalten zulassen. Lediglich der Anteil der Befragten, die das sensitive Merkmal aufweisen, lässt sich ermitteln.

Unsere Auswertungen zeigen, dass alle hier getesteten alternativen Verfahren im Vergleich zur „Sonntagsfrage“ einen signifikant höheren Stimmanteil der AfD ermitteln. Die Konfidenzintervalle des Listen-Experiments sowie der Randomisierten-Antwort-Technik sind jedoch so groß, dass der praktische Nutzen dieser Verfahren zu Prognosezwecken stark eingeschränkt ist. Die Weisheit-der-Vielen-Methode hingegen erreicht nicht nur die akkurateste Punktvorhersage, sondern auch eine moderate Konfidenzintervallgröße und scheint daher gut geeignet zu sein, die traditionelle Berichterstattung hinsichtlich der „Sonntagsfrage“ zu ergänzen. Durch ihr wenig komplexes Design erleichtert diese Methode ebenfalls die akkurate Kommunikation der Ergebnisse. Sie ist daher eine sinnvolle Ergänzung in Vorwahlstudien, die sich ohne großen Aufwand implementieren lässt.

Darüber hinaus haben unsere Ergebnisse offensichtliche Implikationen für die Wahl- und Wählerforschung. Zunächst weisen unsere Ergebnisse darauf hin, dass Wahlvorhersagen vermehrt auf spezialisierte Vorhersageinstrumente zurückgreifen sollten, anstatt (nur) die aktuelle politische Stimmung zu messen und aus ihren Entwicklungen Vorhersagen abzuleiten (Lewis-Beck 2005). Des Weiteren zeigen unsere Ergebnisse, dass sich zukünftige Forschung mehr damit befassen sollte, inwiefern valide Messungen der Wahlabsicht auf der Individualebene ermöglicht werden können. Wir zeigen zwar auf, dass wir Verzerrung anhand von alternativen Methoden sichtbar machen können, opfern dabei aber bewusst die Möglichkeit, Informationen über individuelle Wahlabsichten zu erhalten. Forscherinnen und Forscher, die zum Beispiel die Beweggründe, eine bestimmte Partei zu wählen, untersuchen (etwa mehrere Beiträge in Schmidt-Beck et al. 2014), sind natürlich auf solche Informationen angewiesen. Daher sollten wir unsere Bemühungen verstärken, das Wirken unserer Messinstrumente genauer zu verstehen und bei Möglichkeit zu verbessern.

Auch für die Parteienforschung sind unsere Ergebnisse von großer Bedeutung, da sie Wege skizzieren, Wählerpotentiale und die aktuelle politische Stimmung auch für rechtspopulistische Parteien zu messen. Wir zeigen, dass das Listen-Experiment das Wählerpotential der AfD akkurat bestimmen kann. Ein genaueres Wissen über das Wählerpotential der Parteien

ermöglicht insbesondere Forschung, die einen Zusammenhang zwischen Umfrageergebnissen und Parteiverhalten nachweist (Adams et al. 2015; Bräuninger 2009; Lehmbruch 2000). Sie würde auch davon profitieren, dass die aktuelle politische Stimmung nicht nur durch die „Sonntagsfrage“ gemessen würde, sondern methodologisch breiter aufgestellt wäre. Insbesondere die Weisheit-der-Vielen-Methode erscheint uns eine kostengünstige und einfache Alternative zur Messung politischer Stimmung zu sein.

Des Weiteren können die hier vorgestellten Messmethoden dazu genutzt werden, Veränderungen in der gesellschaftlichen Akzeptanz rechtspopulistischer Parteien sichtbar zu machen, worauf wir in den Schlussfolgerungen weiter eingehen.

Wir beschreiben im nächsten Abschnitt die von uns implementierten alternativen Messmethoden und ihre Logik genauer, bevor wir uns der Auswertung der Ergebnisse zuwenden. Wir schließen mit einem Überblick über unsere Ergebnisse und Implikationen für die weitere Forschung.

2. Alternative Messverfahren

Messmethoden werden in der Regel mit dem Ziel konzipiert, dass sie korrekt die Maßzahl in der Zielpopulation (z.B. den Anteil der AfD-Wähler in der Bevölkerung) erfassen und gleichzeitig zufällige Schwankungen zwischen verschiedenen Messungen minimieren. Messtheoretiker sprechen in diesem Zusammenhang oft vom Signal, das die Forscherinnen und Forscher interessiert, einerseits und von Störgeräuschen, die es herauszufiltern gilt, andererseits. Wenn ein Messinstrument jedoch fehlerhaft das Signal aufnimmt, sind seine Messungen oft nicht brauchbar.[2] In der Umfrageforschung kommt eine weitere Schwierigkeit hinzu: Das Messinstrument muss nicht nur das korrekte Signal aufnehmen, es darf auch nicht der Grund dafür sein, dass Befragte bewusst ein falsches Signal senden. Denn selbst wenn das Messinstrument das Signal perfekt misst, so entspricht seine Messung doch nicht dem, was die Forscherin/der Forscher zu messen glaubt.

Soziale Erwünschtheit führt oftmals dazu, dass Befragte in Umfragen ein verzerrtes Signal senden, also unwahre Antworten oder keine Antwort abgeben. Der Grund dafür ist, dass sie denken, dass die wahre Antwort verpönt sei, also den gesellschaftlichen Erwartungen und Normen nicht entspreche und daher als unerwünscht angesehen wird (Crowne u. Marlowe 1960). Dieser Effekt tritt auch in Online-Befragungen ohne menschlichen Interviewer auf (Kreuter et al. 2008). Dadurch ist im Kontext der Wahlforschung sowohl die Abfrage des aktuellen Stimmanteils rechtspopulistischer Parteien (z.B. Johann et al. 2016; Aichholzer et al. 2014) als auch die Prognosefähigkeit der „Sonntagsfrage“ erheblich eingeschränkt (Schnell u. Noack 2014).

Um die Schwächen der „Sonntagsfrage“ im Hinblick auf sozial erwünschtes Antwortverhalten zu überwinden, testen wir drei alternative Verfahren im Kontext der Wahl zum 19. Deutschen Bundestag und vergleichen sie mit der Vorhersagekraft der „Sonntagsfrage“. Zwei der Verfahren produzieren bewusst Störgeräusche, damit die individuelle Antwort der Befragten geheim bleibt und sie wahrheitsgemäß antworten, also das korrekte Signal senden können. Das dritte Verfahren fragt Umfrageteilnehmer und -teilnehmerinnen nicht direkt nach ihren individuellen Einstellungen, sondern lediglich nach ihrer Einschätzung des Wahlausgangs.

[2] Eine äußerst seltene Ausnahme sind Fälle, in denen die Forscherin jederzeit weiß, wie groß der Fehler ist und ihn deshalb korrigieren kann. Ebenso können auch Veränderungen zwischen fehlerhaften Ergebnissen aussagekräftig sein, wenn die Fehler bei allen Messungen gleichgroß sind.

Wir testen diese Methoden hinsichtlich ihres Nutzens für Wahlprognosen mit Blick auf das Ergebnis der AfD bei der 19. Bundestagswahl. Die AfD gilt inzwischen als rechtspopulistische Partei (z.B. Rosenfelder 2017; Lewandowsky et al. 2016; Lewandowsky 2015) deren Stimmanteil in Umfragen regelmäßig, vermutlich wegen sozialer Unerwünschtheit, unterschätzt worden ist. Dementsprechend gab es bei Landtagswahlen auch wiederholt Schwierigkeiten, das AfD-Ergebnis genau vorherzusagen (Bergmann u. Diermeier 2017). Die Fragen, die wir hier auswerten, wurden im Juli 2017 und somit rund zwei Monate vor der Bundestagswahl im GIP implementiert. Obwohl das GIP eine Online-Befragung ist, stellt das Rekrutierungsverfahren sicher, dass die Befragten repräsentativ die deutsche Bevölkerung widerspiegeln (Blom et al. 2015).

2.1 Randomsierte-Antwort-Technik

Es existieren verschiedene Varianten der Randomisierten-Antwort-Technik (RAT). Sie alle basieren auf der Logik, keinen deterministischen, sondern einen probabilistischen Zusammenhang zwischen der berichteten und der wahren Merkmalsausprägung zu schaffen (Jann et al. 2012). Dazu kombiniert der/die Befragte die Antwort auf eine Ja-Nein-Frage (Signal), bei der „ja“ die sozial unerwünschte Antwort ist, mit Informationen aus einem Zufallsgenerator, dessen Wahrscheinlichkeitsverteilung bekannt ist (Störgeräusche). Da die statistische Verteilung der Störgeräusche bekannt ist, können sie im Aggregat, aber nicht auf Individualebene herausgerechnet werden. Die verschiedenen Varianten der Randomisierten-Antwort-Technik unterscheiden sich darin, welcher Zufallsgenerator genutzt wird und wie er die abgegebene Antwort beeinflusst.

In vielen Anwendungen werfen Befragte zum Beispiel eine Münze und antworten mit „ja“, wenn die Münze auf Kopf fällt, bei Zahl antworten sie wahrheitsgemäß. Bei diesem sogenannten „forced response design“ (Coutts u. Jann 2011) kann nun nicht mehr klar differenziert werden, ob die gegebene Antwort die wahrheitsgetreue Wiedergabe der eigenen Einstellung eines/einer Befragten oder eine erzwungene Antwort ist. Da in etwa die Hälfte aller Befragten „ja“ als Antwort vorgegeben bekommt, lässt sich der ungefähre Anteil der wahrheitsgemäßen „Ja“-Sager auf der Aggregatsebene dennoch leicht berechnen (Boruch 1971). Offensichtlich ist es aber nicht effizient, die wahren Antworten der halben Stichprobe nicht abzufragen. Deswegen entwickelten Forscherinnen und Forscher effizientere Alternativen, um die Antwort der Befragten mit einem Zufallsinstrument zu kombinieren.

In dieser Studie wenden wir ein solches effizienteres RAT-Design, das sogenannte „kreuzweise RAT-Design“, nach Yu, Tian und Tang (2008) an, das sich in anderen Studien bereits als valide erwiesen hat (Höglinger et al. 2016). In diesem RAT-Verfahren werden den Befragten zwei Fragen mit dichotomen Antwortmöglichkeiten gleichzeitig gestellt, die Fragen werden „gekreuzt“. Bei der einen Frage handelt es sich um die sensitive Frage, die das eigentliche Erkenntnisinteresse darstellt. Sie wird mit einer anderen, zusammenhangslosen und nicht sensitiven Frage, bei der die Häufigkeitsverteilung in der Zielpopulation bekannt ist, „gekreuzt“. Der/die Befragte überprüft nun, ob die Antworten auf die beiden Fragen gleich sind und gibt nur diese Information weiter. Die einzelnen Antworten auf die beiden Teilfragen werden nicht direkt abgefragt, wodurch den Befragten glaubhaft versichert wird, dass ihre individuelle Antwort geheim bleibt.

Zunächst präsentieren wir den Befragten einen Einleitungstext, um ihre Bereitschaft zur wahrheitsgemäßen Beantwortung zu erhöhen und um die Methode kurz zu erklären, damit es nicht durch Missverständnisse zu Verzerrungen kommt. Der Einleitungstext lautet:

> *„In unserer Studie probieren wir manchmal auch Neues aus. Wir werden Ihnen gleich zwei Fragen gleichzeitig stellen und Sie bitten, uns zu sagen, ob Ihre Antworten auf beide Fragen gleich sind.*
>
> *Zuerst möchten wir Sie bitten, an einen Ihrer Freunde oder Verwandten zu denken, dessen Hausnummer Sie kennen. Bitte merken Sie sich nun die Zahl, mit der die Hausnummer beginnt und klicken Sie auf 'Weiter'.“*

Die erste Ziffer der Hausnummer eines/einer Bekannten nutzen wir als Zufallsgenerator (Diekmann 2012). Da die Verteilung der ersten Zahl der Hausnummern in Deutschland der Benfordschen Verteilung folgt, beträgt die Wahrscheinlichkeit eine Hausnummer zufällig zu wählen, deren erste Ziffer eine 1, 2, 3 oder 4 ist, 70 Prozent.[3]

Wir nutzen diesen speziellen Zufallsgenerator aus drei Gründen. Erstens ist dieser Zufallsgenerator theoretisch angemessen. Den Befragten wird dadurch glaubhaft versichert, dass die Umfrageforscher die Hausnummer des/der Bekannten, die selbst nicht genannt wird, unmöglich kennen. Das sollte den Befragten die affirmative Antwort von sozial unerwünschten Verhalten erleichtern, da offensichtlich Rückschlüsse auf individuelles Antwortverhalten unmöglich sind. Zweitens lässt sich dieser Zufallsgenerator im Rahmen einer Online-Umfrage schnell und einfach umsetzen, da er keinen physischen Gegenstand wie etwa eine Münze benötigt.

Drittens erhöht der Zufallsgenerator die Effizienz des Schätzers, da es eine Diskrepanz zwischen subjektiv wahrgenommener und objektiver Wahrscheinlichkeit gibt, die als Benfordsche Illusion bekannt ist (z.B. Diekmann 2012). Theoretisch beruht die effizienteste Implementierung dieses Designs auf einer zweiten (nicht-sensitiven) Frage, der jeder/jede Befragte zustimmt, da beide Antworten nur dann gleich sind, wenn Befragte auch der sensiblen Frage zustimmen.[4] Offensichtlich führt diese Implementierung jedoch dazu, dass die Anonymität der Antworten nicht gewährleistet ist, wodurch das Problem der sozialen Erwünschtheit erneut das Signal verzerrt. Ein Münzwurf mit der gleichen Chance, Kopf oder Zahl zu erhalten, stellt die ineffizienteste, aber anonymste Implementierung dar. Die Benfordsche Illusion erlaubt uns, einen Kompromiss zwischen Anonymität und Effizienz zu erhalten, weil die objektive Wahrscheinlichkeit sich von der subjektiv empfundenen Wahrscheinlichkeit unterscheidet. Befragte unterschätzen in der Regel die Häufigkeit kleiner Zahlen und überschätzen die Häufigkeit großer Zahlen. Somit erhöht die Benfordsche Verteilung als Zufallsgenerator die Effizienz des Schätzers, was insbesondere mit Blick auf Wahlprognosen wichtig ist (Diekmann 2012).

Die Befragten beenden die Anzeige des Einleitungstexts durch einen Klick auf die Schaltfläche 'Weiter'. Nun erscheint der eigentliche Fragetext. Die Befragten werden ausdrücklich darauf hingewiesen, dass sie nicht die Fragen beantworten, sondern lediglich angeben sollen, ob die Antworten identisch sind. Der Wortlaut der Frage ist:

[3] Zur Überprüfung dieser Annahme haben wir die Verteilung der ersten Ziffern der Hausnummern aller GIP Befragten ausgewertet. Da es sich um eine repräsentative Stichprobe handelt, zeigt diese Auswertung, dass die Verteilungsannahme gerechtfertigt ist und dass circa 70 % der Hausnummern in Deutschland mit Ziffern zwischen 1 und 4 beginnen. Abbildung A2 im Anhang verdeutlicht dies.

[4] Selbstverständlich könnte es auch eine Frage sein, der niemand zustimmt.

„Bitte geben Sie an, ob Ihre Antworten auf diese beiden Fragen gleich sind, also beide Antworten entweder ‚ja' oder ‚nein', oder unterschiedlich sind, also eine Antwort ‚ja' und eine Antwort ‚nein'.

- *Werden Sie bei der kommenden Bundestagswahl mit Ihrer Zweitstimme die Partei Alternative für Deutschland (AfD) wählen?*
- *Wir haben Sie gerade gebeten, an einen Ihrer Freunde oder Verwandten zu denken, dessen Hausnummer Sie kennen. Ist die Zahl, mit der die Hausnummer beginnt, eine 1, 2, 3 oder 4?"*

Die Befragten können mit „gleich (beide Antworten entweder 'ja' oder 'nein')" oder „ungleich (eine Antwort 'ja' und eine Antwort 'nein')" antworten. Die Reihenfolge der zwei Teilfragen ist randomisiert, um eine mögliche Verzerrung durch Listeneffekte zu verhindern.

Yu et al. (2008, S. 256) zeigen, dass in diesem Kontext der Anteil der AfD-Wähler wie folgt berechnet werden kann:

$$\hat{\pi} = \left(\frac{r}{n} + p - 1\right) / (2p - 1)$$

wobei $\hat{\pi}$ der geschätzte Anteil der AfD-Wähler und Wählerinnen in der Bevölkerung ist, *r* ist die Anzahl der Befragten, die angeben, dass die Antwort auf beide Fragen gleich sind, *n* ist die Anzahl der Befragten, die die Frage beantworten und *p* ist die Wahrscheinlichkeit, dass die Hausnummer mit 1, 2, 3 oder 4 beginnt (p=0,7).[5]

Allerdings hat dieses kreuzweise RAT-Design einige Schwachstellen, die die Nützlichkeit dieser Methode in unterschiedlichen Kontexten beeinträchtigen. So müssen die Befragten das Verfahren verstehen und erkennen, dass es ihnen Anonymität gewährt. Zudem müssen sie bereit sein, den Anweisungen zu folgen (Jann et al. 2012; Krumpal 2012; Coutts u. Jann 2011; Yu et al. 2008). Dies ist nicht immer gewährleistet (Coutts u. Jann 2011; Höglinger et al. 2016). Coutts und Jann (2011) kommen daher zu dem Fazit, dass das Listen-Experiment eine überlegene Alternative zum RAT-Design darstellt. Um diese Schlussfolgerung vor dem Hintergrund der Vorhersage von Wahlergebnissen rechtspopulistischer Parteien zu testen, implementieren wir in der Studie ebenfalls eine Variante des Listen-Experiments.

2.2 Listen-Experiment

Bei der klassischen Variante des Listen-Experiments handelt es sich um ein Design, das darauf beruht, die Befragten zufällig in zwei Experimentalgruppen einzuteilen. Beide Gruppen erhalten dann eine Liste mit verschiedenen nicht-sensitiven Elementen. Der einzige Unterschied zwischen den Gruppen besteht darin, dass bei einer Gruppe (der Treatmentgruppe) diese Liste um das sensitive Element erweitert wird. Die Befragten werden dann darum gebeten anzugeben, *wie viele* (und eben nicht *welche*) der aufgelisteten Elemente auf sie zutreffen. Für die Befragten ist es daher einfach zu erkennen, dass ihre Antwort keinerlei Rückschlüsse auf ihr individuelles Verhalten hinsichtlich des sensitiven Elements zulässt, wodurch ihr Antwortverhalten nicht durch das Problem der sozialen Erwünschtheit beeinflusst wird (Coutts u. Jann 2011). Auch ist die Auswertung der Ergebnisse wenig kompliziert. So wird der Anteil derer, die das sensitive Verhalten aufweisen, durch die Differenz der durchschnittlichen Anzahl an Verhaltensweisen zwischen der Treatment- und der Kontrollgruppe ermittelt (z.B. Rosenfeld et al. 2016; Glynn 2013).

[5] Yu et al. (2008) bestimmen auch die Variabilität des Schätzers. Aus unten genannten Gründen verwenden wir aber die durch Bootstrapping bestimmte Variabilität.

Genau wie das oben vorgestellte RAT-Design ist das Listen-Experiment jedoch im Vergleich zur direkten Frage ineffizient. Allerdings kommt die Ineffizienz nicht wie beim RAT-Design von einer zusätzlichen Quelle zufälliger Abweichungen, sondern (i) von der Aufteilung in Kontroll- und Treatmentgruppe, die dazu führt, dass nur die Hälfte der Befragten das sensitive Element beantworten, und (ii) von der Aggregation von Informationen. Die Befragten evaluieren nicht einzeln die Elemente der Liste, sondern geben lediglich die aggregierte Anzahl an affirmativen Antworten an (Rosenfeld et al. 2016).

Johann et al. (2016) haben eine Variante dieser Methodik bereits im Kontext der Wahl zum 18. Deutschen Bundestag 2013 getestet, um den Stimmanteil der AfD zu ermitteln. Basierend auf einer nichtrepräsentativen Online-Studie kommen sie zu dem Ergebnis, dass das Listen-Experiment wenig geeignet ist, um die Probleme der „Sonntagsfrage“ zu lösen. Trotz dieses ernüchternden Fazits bleibt offen, ob die Ergebnisse auf die Schwächen der Methodik zurückzuführen oder ein Artefakt der nichtrepräsentativen Stichprobenziehung sind. Zudem testen die Autoren die klassische Implementierung des Listen-Experiments, bei der eine Liste erstellt und Befragte entweder in der Kontroll- oder in der Treatmentgruppe sind (Kuklinski et al. 1997).

Im Gegensatz dazu implementieren wir ein *doppeltes* Listen-Experiment in einem repräsentativen Online-Survey (Glynn 2013; Droitcour et al. 1991). Den Befragten werden jeweils zwei unterschiedliche Listen mit Elementen vorgelegt, jeweils eine mit und eine ohne das sensitive Element, und sie werden gebeten die Anzahl der auf sie zutreffenden Elemente anzugeben. Beide Gruppen dienen somit sowohl als Kontroll- wie auch als Treatmentgruppe. So bekommt Experimentalgruppe 1 die Liste A ohne das sensitive Element und Liste B mit diesem Element vorgelegt. Experimentalgruppe 2 erhält Liste A mit dem sensitiven Element und Liste B ohne. Gleichzeitig kann Experimentalgruppe 1 als Kontrollgruppe für Experimentalgruppe 2 hinsichtlich der Liste A fungieren und umgekehrt. Die Präzision der Schätzer wird durch die Implementierung des doppelten Listen-Experiments erhöht, da jeder Befragte nach seiner AfD-Wahlabsicht befragt wird (Coutts u. Jann 2011; Droitcour et al. 1991). Nichtsdestotrotz bleiben die Störsignale, die aus der Aggregation mehrerer Antworten in der Liste entstehen, erhalten.

Da die Befragung im Juli 2017 - und somit circa zwei Monate vor der Bundestagswahl - durchgeführt wurde, werden die Befragten gebeten anzugeben, wie viele der Punkte auf der Liste sie wahrscheinlich in den nächsten drei Monaten tun werden.[6] Tabelle 1 zeigt die Elemente der Listen A und B, wobei jeweils das vierte Element das sensitive Item (Treatment) ist. Experimentalgruppe 1 erhält Liste A ohne dieses Element und Liste B mit Element 4. Folglich erhält Experimentalgruppe 2 Liste A mit und Liste B ohne Element 4. Um die Effizienz des Schätzers zu erhöhen, sind die Elemente der Listen so gestaltet, dass es eine möglichst hohe positive Korrelation zwischen den Listen und eine negative Korrelation mancher Items innerhalb der Listen gibt (Glynn 2013). Schließlich begegnen wir möglichen „ceiling effects" und „floor effects", d. h., dass Befragte jedes oder keins der Elemente der Liste tun werden und wir somit doch auf individuelles Verhalten schließen können, indem zum einen Elemente in die Liste aufgenommen wurden, die (nahezu) jeder Befragte bejaht und zum anderen Elemente, die (nahezu) jeder Befragte verneint (z.B. Johann et al. 2016; Rosenfeld et al. 2016; Glynn 2013; Kuklinski et al. 1997).

[6] Der genaue Wortlaut der Frage ist: "Wie viele dieser Punkte werden Sie wahrscheinlich in den nächsten drei Monaten tun?".

Tabelle 1: Fragelisten des Listen-Experiments

	Liste A	Liste B
1	ein Wahlplakat genauer anschauen	mich mit Freunden oder Familienmitgliedern über Politik unterhalten
2	den Wetterbericht anschauen	die Tagesschau anschauen
3	mich an einer Demonstration beteiligen	mich ehrenamtlich engagieren
4	*bei der kommenden Bundestagswahl die Alternative für Deutschland (AfD) wählen*	*bei der kommenden Bundestagswahl die Alternative für Deutschland (AfD) wählen*
5	die Wahlprogramme aller aktuell im Deutschen Bundestag vertretener Parteien lesen	mich als Kandidat für ein politisches Amt zur Wahl stellen

Nach Droitcour et al. (1991, S. 189) berechnet sich dann der Anteil der AfD-Wähler und Wählerinnen in der Bevölkerung als

$$\hat{\pi} = \frac{(\bar{X}_{A5} - \bar{X}_{A4}) + (\bar{X}_{B5} - \bar{X}_{B4})}{2}$$

wobei $\bar{X}_{A5}$ der Mittelwert der Antworten ist, die auf Liste A (erster Index) mit fünf Antworten (zweiter Index) abgegeben wurden.[7]

2.3 Weisheit-der-Vielen Methode

Verglichen mit den anderen beiden hier vorgestellten Methoden, ändert die Weisheit-der-Vielen (WdV) Methode die Perspektive der Frage (Murr 2017). Anstatt Befragte nach dem sensitiven Merkmal zu fragen, werden sie lediglich gebeten einzuschätzen, wie hoch der Anteil der Personen ist, die das sensitive Merkmal aufweisen. Im Kontext der Bundestagswahl 2017 fragen wir die GIP-Teilnehmer und Teilnehmerinnen daher:

> „*Was denken Sie, wie viel Prozent der Zweitstimmen wird die Partei Alternative für Deutschland (AfD) bei der kommenden Bundestagswahl im September wohl bekommen? Die Zweitstimme ist die Stimme, mit der eine Partei gewählt wird*."

Die Befragten werden gebeten, in einem Antwortfeld ihre Prognose in Form einer Zahl zwischen 0 und 100 einzugeben. Die Prognose ergibt sich aus dem arithmetischen Mittel der individuellen Antworten.[8]

Sicherlich ist damit zu rechnen, dass viele Befragte sich deutlich verschätzen, denn eine Einschätzung über sich selbst, wie sie die anderen Fragentypen vorsehen, ist deutlich einfacher vorzunehmen als eine Prognose des Verhaltens von Millionen anderer Menschen. Zentrale Ergebnisse der Sozialwahltheorie zeigen aber, dass unter sehr generellen Bedingungen trotzdem eine qualitativ hochwertige Vorhersage möglich ist, die „Weisheit der Vielen" eben.

Ausgangspunkt ist dabei das Condorcet-Jury-Theorem (Condorcet 1785), das zeigt, dass unter recht einfachen Bedingungen die aggregierte Meinung einer Gruppe deutlich genauer ist als die Meinung von Individuen (z.B. Condorcet 1785; Galton 1907). Das berühmte Theorem besagt, dass die Wahrscheinlichkeit einer Gruppe die richtige Entscheidung zwischen zwei

[7] Auch Droitcour et al. (1991) bestimmen die Variabilität des Schätzers analytisch. Im Gegensatz dazu nutzen wir jedoch Bootstrap-Standardfehler, um die Vergleichbarkeit der getesteten Methoden zu wahren.
[8] Die Variabilität des Schätzers ist deshalb die übliche Variabilität des Mittelwertes.

Alternativen mittels Mehrheitsbeschluss zu treffen, rapide mit der Größe der Gruppe ansteigt, wenn die individuelle Wahrscheinlichkeit jedes Gruppenmitglieds die richtige Entscheidung zu treffen größer als 50 Prozent ist. Diverse anschließende Arbeiten generalisieren dieses Theorem und zeigen, dass der Mechanismus auch in Situationen mit heterogenen individuellen Wahrscheinlichkeiten, korreliertem Abstimmungsverhalten und mehreren Wahlalternativen Bestand hat (z.B. List u. Goodin 2001; Boland 1989; Grofman et al. 1983).

Dadurch, dass die Weisheit-der-Vielen-Methode keine sensitive Frage bezüglich des eigenen Verhaltens stellt und deswegen auch keine zusätzlichen Störsignale benötigt, um eine Anonymität der individuellen Antworten zu gewährleisten, hat sie gegenüber den anderen vorgestellten Methoden einen entscheidenden Vorteil: Die Effizienz des Schätzers und die Anonymität der Befragten stehen nicht im Konflikt miteinander. Vielmehr ist die vergleichbar hohe Effizienz der Methode ein großer Vorteil, denn sie ermöglicht präzise Prognosen ohne, wie die „Sonntagsfrage", Gefahr zu laufen, ein verzerrtes Signal zu erzeugen (Graefe 2014; Galton 1907). Während RAT-Design und Listen-Experiment den Befragten das Senden des richtigen Signals durch Anonymisierung erleichtern, aber ineffizient sind, misst die „Sonntagsfrage" ein systematisch verzerrtes Signal effizient. Im Gegensatz dazu erzeugt die Weisheit-der-Vielen-Methode keinen Widerspruch zwischen Anonymität und Effizienz.

3. Empirischer Vergleich der Messmethoden

3.1 Methodologische Vorbemerkungen zum Vergleich der Methoden

Die oben beschriebenen Methoden zur Vorhersage des Stimmanteils der AfD wurden rund zwei Monate vor der Bundestagswahl am 24.09.2017 in der 30. Welle der GIP-Befragungen im Juli 2017 implementiert. Um einer Verzerrung der Ergebnisse durch das mehrfache Abfragen desselben sensitiven Items vorzubeugen, wurden die Befragten zunächst zufällig drei gleich großen Gruppen zugewiesen. Die „Sonntagsfrage" und die Weisheit-der-Vielen-Frage wurden allen Befragten gestellt (siehe Tabelle 2).[9] Ein zufällig ausgewähltes Drittel der Befragten wurde der RAT-Befragung zugeordnet, das zweite zufällig ausgewählte Drittel erhielt das Listen-Experiment mit Liste A als Treatment und Liste B als Kontrollliste und das letzte Drittel der Befragten erhielt ebenfalls das Listen-Experiment, jedoch mit Liste B als Treatment und Liste A als Kontrollliste.[10] Aus dieser zufälligen Zuteilung ergeben sich Folgen für unsere Analysestrategie, die wir nun ausführlicher beschreiben.

Tabelle 2: Fragentyp und fehlende Antworten

Fragentyp	Befragte	Anteil Befragte in Prozent	Fehlende Antworten	Verwendbare Antworten in Prozent
Sonntagsfrage	2597	100,0	481	81,5
WdV	2597	100,0	32	99,0
RAT	867	33,4	94	89,1
Listenexperiment	1730	66,6	7	99,6
davon Liste A	865	33,3	3	99,7
davon Liste B	865	33,3	4	99,5

[9] Wir stellen in der Tabelle nur Befragte dar, die die "Sonntagsfrage", die WdV-Frage sowie die RAT-Fragen oder beide ICT-Fragen gestellt bekommen haben. 17 Befragte beantworteten die Sonntagfrage, stiegen dann aber aus der Umfrage aus, bevor die anderen Fragen gestellt werden konnten.

[10] In Anhang A1 zeigen wir, dass die zufällige Aufteilung erreicht, dass die Befragtengruppen in wesentlichen anderen Merkmalen gleich sind und daher Unterschiede zwischen den Gruppen auf die Gruppenzugehörigkeit (Fragetyp) zurückzuführen zu sind.

Wir möchten herausfinden, wie erfolgreich die verschiedenen Methoden darin sind, das AfD-Wahlergebnis bei der Bundestagswahl 2017 schon etwa zwei Monate vor der Wahl zu prognostizieren. Dazu vergleichen wir, wie nah die Vorhersagen dem tatsächlichen AfD-Ergebnis kommen und inwiefern die Vorhersagen einer Methode streuen. Die Streuung messen wir anhand von Konfidenzintervallen, die kleiner werden, wenn mehr Daten zu Verfügung stehen. Es wäre also unangemessen, den gesamten Datensatz der GIP-Umfrage zu nutzen, um die Streuungen der „Sonntagsfrage“ und des RAT-Schätzers miteinander zu vergleichen, da drei Mal so viele Datenpunkte für die „Sonntagsfrage“ zur Verfügung stehen und die Konfidenzintervalle deswegen kleiner sein sollten.

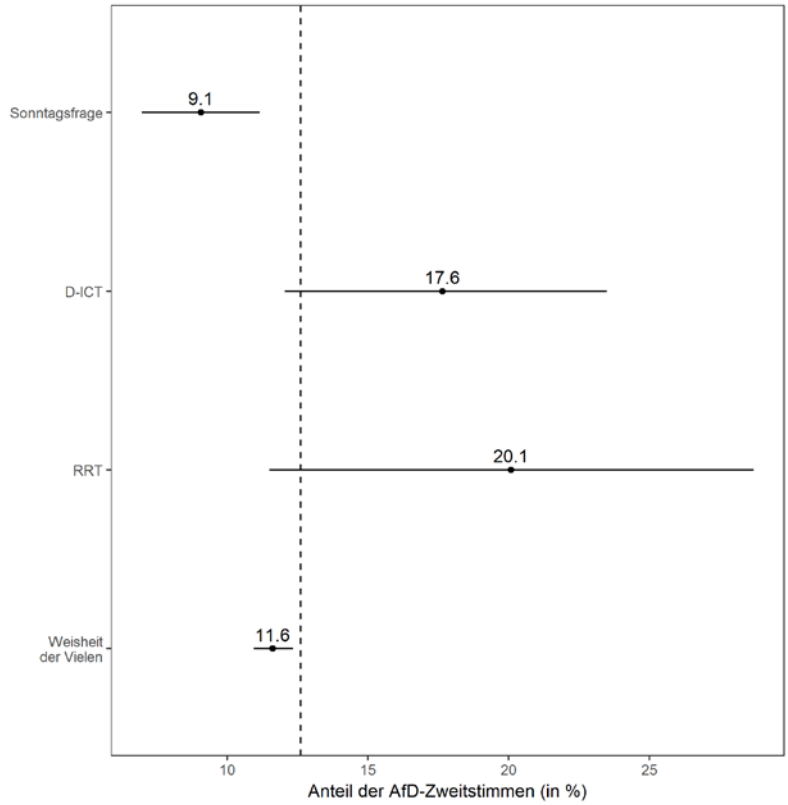
Im Folgenden wenden wir deshalb eine Art Bootstrapping an, um den Datensatz der „Sonntagsfrage“ sowie der Weisheit-der-Vielen-Frage zu „verkleinern“ und angemessene Vergleiche zwischen den Methoden vornehmen zu können. Konkret ziehen wir mit Zurücklegen aus dem Gesamtdatensatz mit 2597 Befragten 865 Befragte zufällig heraus und berechnen den Anteil der AfD-Wähler in diesem Teildatensatz. Dabei werden auch Befragte berücksichtigt, die keine Antwort abgeben wollen oder angeben, nicht wahlberechtigt zu sein, denn wir sind auch daran interessiert, wie die Effektivität der Methoden sinkt, weil Befragte sich weigern, Fragen zu beantworten oder sie nicht verstehen. Dieses Resampling führen wir 1000 Mal durch, schätzen jeweils den Anteil der AfD-Wähler und Wählerinnen und nutzen das 2,5. bzw. 97,5. Perzentil der resultierenden Verteilung der AfD Anteile als Grenzen des 95 Prozent-Konfidenzintervalls. Den Mittelwert dieser Schätzungen verwenden wir als Punktschätzer. Wir nutzen dieselbe Methode auch für die Weisheit-der-Vielen-Frage. Als Ergebnis erhalten wir Schätzungen und die dazugehörigen 95 Prozent-Konfidenzintervalle, die mit den Ergebnissen der Methoden vergleichbar sind, die nur 865 Befragte zur Verfügung haben. Um Vergleichbarkeit zwischen allen Designs herzustellen, berechnen wir die Variabilität aller Schätzer durch einen Bootstrap.

3.2 Ergebnisse

Abbildung 1 stellt eine Übersicht der Ergebnisse dar. Die Punkte in der Grafik sind die ermittelten Punktschätzer des AfD-Anteils an den Zweistimmen, horizontale Linien die entsprechenden 95 Prozent-Bootstrapping-Konfidenzintervalle. Die vertikale, gestrichelte Linie zeigt das wahre Zweitstimmenergebnis der AfD bei der Bundestagswahl an (12,6 Prozent).
Als erstes fallen die großen Schwankungen zwischen den verschiedenen Messmethoden auf – sowohl in Bezug auf deren Nähe zum wahren Ergebnis, als auch auf deren Variabilität. Die „Sonntagsfrage“ unterschätzt die AfD-Stimmenanteile um gut 3,5 Prozentpunkte, was 28 Prozent des tatsächlich realisierten AfD-Zweitstimmenergebnisses (gestrichelte Linie) entspricht. Die Spannweite des Konfidenzintervalls beträgt 4,2 Prozentpunkte. Im Vergleich dazu lag die AfD in Umfragen im Juli zwischen 7 Prozent (Allensbach) und 9 Prozent (Infratest Dimap)[11], wie auch bei Vorhersageportalen wie etwa Zweitstimme.org (Munzert et al. 2017).

Abbildung 1: Erwarteter AfD-Zweitstimmenanteil bei der Bundestagswahl 2017

[11] Publizierte Umfragen werden freundlicherweise vom Internetportal wahlrecht.de gesammelt und aufbereitet (http://www.wahlrecht.de/umfragen/).



Anmerkungen: Die Punkte stellen Punktschätzungen dar, die durchgezogenen Linien die entsprechenden 95 Prozent-Konfidenzintervalle. Die gestrichelte Linie stellt das tatsächliche AfD-Zweitstimmenergebnis dar. Quelle: GIP Welle 30.

Es gibt im Wesentlichen drei Gründe, die zu dieser Abweichung der Umfrageergebnisse zum amtlichen Wahlergebnis führen können. Als erstes kann soziale Erwünschtheit Befragte davon abhalten, ihre AfD-Wahlabsicht zu äußern.[12] Zweitens können Befragte zwischen Befragung und Stimmabgabe ihre Meinung ändern. Das kann prinzipiell in zwei unterschiedliche Richtungen gehen. Befragte können sich bis zum Wahltag der AfD zuwenden oder, falls zuvor schon bei der „Sonntagsfrage" als AfD-Wähler/Wählerin identifiziert, am Wahltag doch noch für eine andere Partei stimmen. [13] Schließlich können Stichprobeneffekte eine Rolle spielen, etwa wenn in der GIP-Stichprobe AfD-Wähler/Wählerinnen zufällig unterrepräsentiert sind. Wir können anhand unserer Daten nicht herausfinden, inwiefern die verschiedenen Fehlerquellen Einfluss auf den Vorhersagefehler der „Sonntagsfrage" nehmen. Es ist aber wahrscheinlich, dass sie alle eine Rolle spielen. Wir wenden uns nun der Frage zu, ob andere Methoden bessere Vorhersagen zulassen als die klassische „Sonntagsfrage".

Das doppelte Listen-Experiment überschätzt das AfD-Ergebnis um 5,4 Prozentpunkte. Sein breites Konfidenzintervall, das 11,5 Prozentpunkte weit ist, schließt aber das wahre AfD-Ergebnis bei der Bundestagswahl ein. Ähnlich verhält es sich bei der RAT. Zwar überschätzt auch diese Messmethode das AfD-Ergebnis deutlich um 7,5 Prozentpunkte, was auf ein in diesem Zusammenhang bisher nur wenig beachtetes Problem von falsch positiven Werten hindeutet (z.B. Höglinger u. Diekmann 2017), aber ihr sogar noch breiteres Konfidenzintervall (mehr als 17 Prozentpunkte) schließt das wahre Ergebnis ein. Statistisch gesprochen wäre es nun einfach, Listenexperiment und RAT als besseres Vorhersaginstrument als die „Sonn-

[12] Knapp 5 % der Befragten verweigern auf die „Sonntagsfrage" die Antwort, weil sie ihre Wahlintention nicht preisgeben wollen. Einige Befragte tun dies vermutlich auch aufgrund sozialer Erwünschtheit.

[13] 13,7 % der Befragten antworten auf die „Sonntagsfrage" mit „weiß nicht", wozu auch unentschlossene Wähler gehören können.

tagsfrage“ zu bezeichnen, da ihre Schätzer nicht signifikant vom Wahlergebnis abweichen und beide Methoden einen höheren Anteil an sozial unerwünschtem Verhalten zu Tage befördern. In der Praxis sind Messungen mit derart hohen Schwankungsbreiten aber nicht nützlich. So umschließen die Grenzen der Konfidenzintervalle sowohl den Fall, dass die AfD mit Grünen, Linken und der FDP um den dritten Platz im Parlament streitet, als auch das Szenario, dass sie die SPD deutlich überholt und sogar in die Nähe des Union-Ergebnisses vorstoßen kann. Im Kontext der Bundestagswahl scheinen die zusätzlichen Störgeräusche, die RAT und doppeltes Listen-Experiment produzieren, um validere Messungen zu ermöglichen, so stark zu sein, dass sie viele interessante Rückschlüsse nicht zulassen.

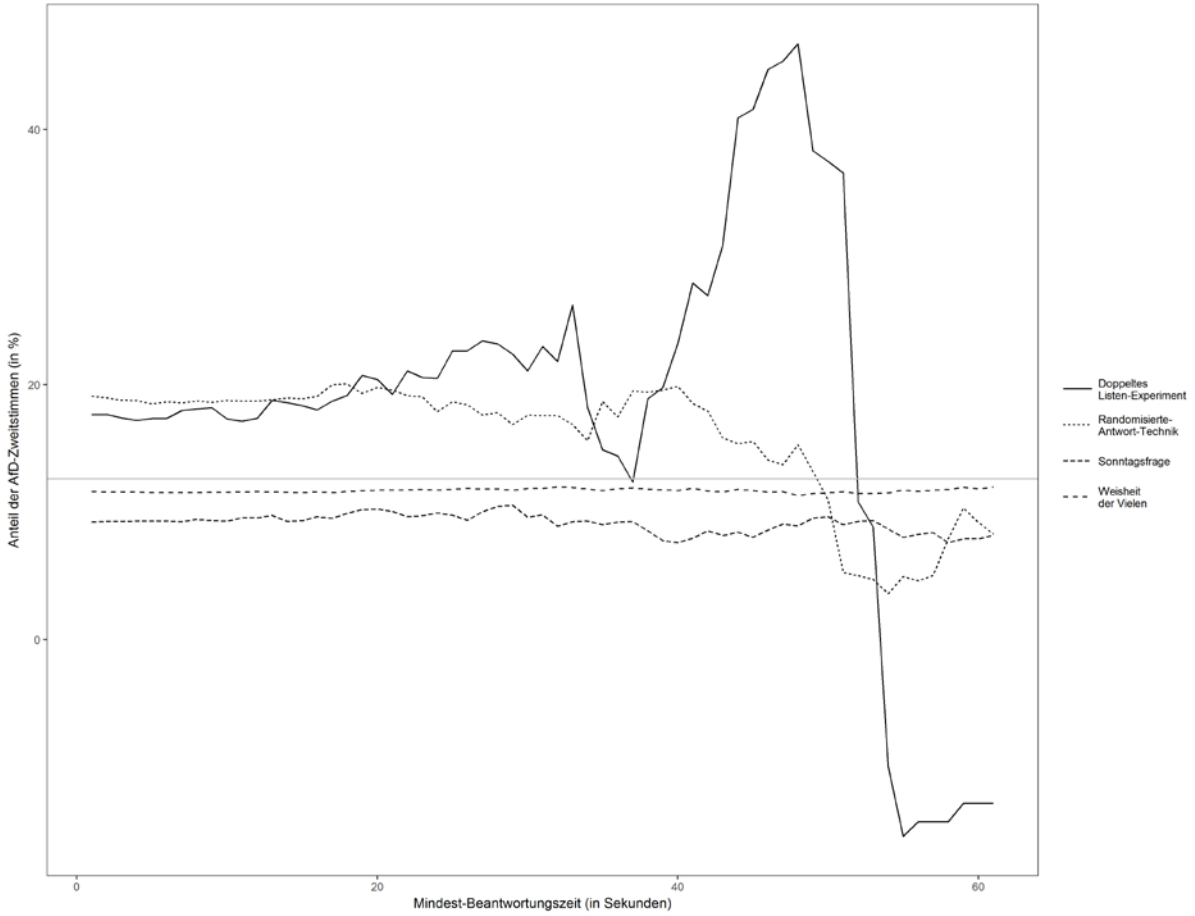
Die Weisheit-der-Vielen-Methode unterschätzt die AfD-Stimmanteile gerade einmal um einen Prozentpunkt, während das Konfidenzintervall mit einer Spannweite von 1,4 Prozentpunkten deutlich kleiner ist als bei allen anderen Schätzern. Der tatsächliche AfD-Stimmanteil bei der Bundestagswahl ist allerdings nicht im Konfidenzintervall eingeschlossen.
Es ist durchaus denkbar, dass die Wählerinnen und Wähler nach der GIP-Befragung im vergangenen Juli ihre Wahlabsicht bis zum Wahltag noch in beide Richtungen ändern. Sie könnten zur AfD umschwenken oder eine bereits getroffene Wahlabsicht wieder revidieren. Daher ist auch nicht zu erwarten, dass der in der Stichprobe durch die „Sonntagsfrage“ ermittelte Wert dem Wahlergebnis entspricht. Tatsächlich zeigen die Antworten der Befragten auf die Weisheit-der-Vielen-Methode, welche explizit auf den prognostizierten Stimmenanteil der AfD am Wahltag abhebt, dass im Schnitt die Befragten einen höheren Stimmanteil erwarten, als durch die „Sonntagsfrage“ zu erwarten wäre. Interessanterweise ist der von der Weisheit-der-Vielen-Methode ermittelte Wert auch höher als der durchschnittlich ermittelte Wert der Umfragen oder der Prognosen von Vorhersagemodellen, die zu dem Zeitpunkt, in dem die GIP-Befragung stattfand, erhoben wurde. Somit ist die Aussagekraft der Weisheit-der-Vielen-Methoden bemerkenswert. Vor allem vor dem Hintergrund, dass das Konfidenzintervall anzeigt, dass mehrfache Replikationen der Frage in anderen Stichproben nur kleine Veränderungen in den Ergebnissen erwarten lassen würden, sodass sich ein klares Bild ergibt, das medial leicht zu berichten ist.

4. Validierungs- und Sensitivitätsanalyse

4.1 Analyse der Antwortzeiten

Sowohl die RAT-Methode als auch das doppelte Listen-Experiment sind kognitiv anspruchsvoller als die direkten Fragen („Sonntagsfrage“ und Weisheit-der-Vielen-Frage), da Befragte nicht wie gewohnt eine einzige Frage beantworten können, sondern mehrere Fragen evaluieren, die Antworten aggregieren und dann diese aggregierte Antwort wiedergeben müssen. Es ist deshalb möglich, dass Befragte vor der Aufgabe oder dem vergleichsweise langen Einleitungstext zurückschrecken und die Aufgabe nicht gründlich bearbeiten (satisficing), um den Fragebogen möglichst bald abzuschließen (Krosnick 1991, 1999). So finden sich insbesondere hinsichtlich des RAT-Designs in der Literatur Belege dafür, dass Befragte die gegebenen Anweisungen nicht befolgen (z.B. Coutts u. Jann 2011). In diesem Fall ist es möglich, dass schnell abgegebene Antworten, die auf eine äußerst kurze Bearbeitungszeit schließen lassen, nicht denselben Informationsgehalt haben wie langsamere Antworten. Da das GIP-Team Daten zur Bearbeitungszeit der Befragten zur Verfügung stellt, testen wir, ob die Ergebnisse der Methoden sich verändern, wenn schnelle Antworten ignoriert werden.

Abbildung 2: Erwarteter AfD-Zweitstimmenanteil bei der Bundestagswahl 2017 nach Antwortzeit



Anmerkungen: Die schwarzen Linien stellen Punktschätzer dar, die graue Linie das tatsächliche AfD-Zweitstimmenergebnis. Quelle: GIP Welle 30.

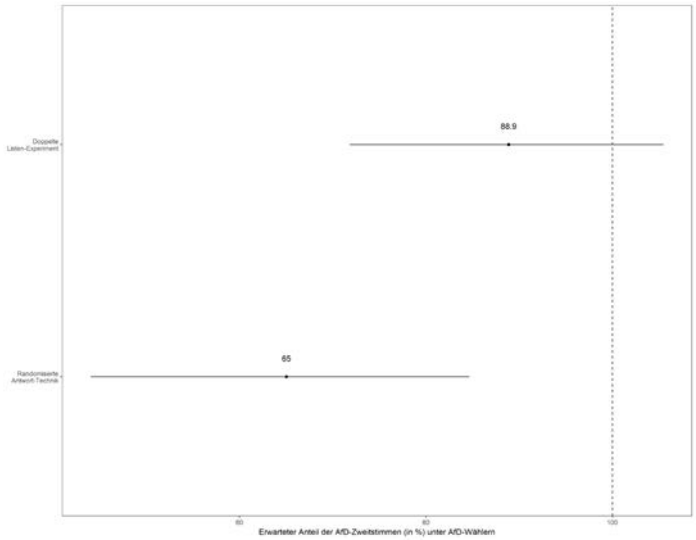
Abbildung 2 zeigt, wie sich die Vorhersagen der verschiedenen Methoden entwickeln, wenn man Antworten ignoriert, die in weniger als einer bestimmten Zeit abgegeben wurden (x-Achse). Am linken Ende des Graphen (Mindest-Beantwortungszeit = 0 Sekunden) entsprechen die Vorhersagen den Prognosen aus Abbildung 1. Mit zunehmender Zeit fallen mehr und mehr Beobachtungen aus der Vorhersage heraus.

Es wird schnell sichtbar, dass mit zunehmendem Wegfall von Beobachtungen Schwankungen in den Vorhersagen auftreten. Davon sind insbesondere das doppelte Listen-Experiment und die RAT betroffen, deren Prognosen dadurch deutlich ungenauer werden. Die „Sonntagsfrage“ sowie die Weisheit-der-Vielen-Frage hingegen, und das ist nicht unbedingt zu erwarten gewesen, sind recht stabil in ihren Vorhersagen, auch wenn viele Befragte aus der Stichprobe herausfallen. Wesentliche Schwankungen treten bei Prognosen auf Basis der „Sonntagsfrage“-erst nach etwa 30 Sekunden Beantwortungszeit auf, da sich dort der Datensatz auf weniger als ein Viertel seiner ursprünglichen Größe verringert. Die Anzahl der genutzten Beobachtungen nimmt für die Weisheit-der-Vielen-Frage ähnlich schnell ab, dennoch bleiben die Vorhersagen stabil und näher am wahren AfD-Ergebnis als alle anderen Vorhersagen. Wir schließen aus diesen Beobachtungen, dass sich die hier gefundenen Unterschiede zwischen den Methoden auch nicht dadurch verändern, dass gegebenenfalls übereilte Antworten ignoriert werden.

4.2 Validierung auf Basis von AfD-Wähler und Wählerinnen

Wir testen die Validität der Ergebnisse auch auf einem weiteren Weg. Dazu betrachten wir nur Befragte, die bei der „Sonntagsfrage“ angeben, AfD wählen zu wollen. Unter der Annahme, dass all diese Befragten auch im Listen-Experiment und der RAT angeben sollten AfD wählen zu wollen, sollten die entsprechenden Vorhersagen einen AfD-Zweitstimmenanteil von 100 Prozent vorhersagen.[14] Abgesehen davon, dass wir die Daten nach der abhängigen Variablen auswählen, bleiben die Analysen unverändert. Abbildung 3 zeigt die Ergebnisse.

Abbildung 3: Erwarteter AfD-Zweitstimmenanteil bei der Bundestagswahl 2017 unter AfD-Wählern und Wählerinnen



Anmerkungen: Die Punkte stellen Punktschätzer dar, die durchgezogenen Linien die entsprechenden 95 Prozent-Konfidenzintervalle. Die gestrichelte Linie stellt das theoretisch erwartete AfD-Zweitstimmenergebnis dar.
Quelle: GIP Welle 30.

Das doppelte Listen-Experiment unterschätzt den AfD-Zweitstimmenanteil um 11 Prozentpunkte, doch das 34 Prozentpunkte weite 95 Prozent-Konfidenzintervall schließt die 100 Prozent-Marke ein. Das RAT-Verfahren verschätzt sich sogar um 35 Prozentpunkte und das Konfidenzintervall beginnt erst 15 Prozentpunkte unter dem erwarteten, „wahren“ Ergebnis. Diese Ergebnisse zeigen, dass im Kontext der Wahlumfragen beide Verfahren zwar das Problem der sozialen Erwünschtheit und die damit verbundene Unterschätzung des AfD-Stimmenanteils korrigieren, sie gleichzeitig jedoch eine Vielzahl zusätzlicher Probleme wie Ineffizienz der Schätzer, erhöhte kognitive Anforderungen an Befragte und die Möglichkeit der Nicht-Kooperation durch Befragte mit sich bringen. Das Hinzufügen von Störgeräuschen scheint trotz dessen, dass es in der Theorie eine elegante Lösung darstellt, in der Praxis nicht zu funktionieren.

[14] Die Wahlabsichtsfrage ist ein Instrument basierend auf nur einer Frage. Die Analyse nimmt an, dass wir diese Wahlabsicht ohne Messfehler messen können. Denkbar ist aber auch, dass sich Befragte bei der Wahlabsichtsfrage unsicher sind und daher inkonsistente Antworten geben, wenn sie erneut zu einer Wahlabsicht befragt werden. Außerdem könnte diese Unsicherheit auch dazu führen, dass, obwohl Befragte angeben AfD zu wählen, sie dies aber am Wahltag nicht tun.

4.3 Validierung des gemessenen AfD-Wählerpotentials

Die Analysen unter AfD-Wählern und Wählerinnen zeigen, dass sowohl das Listen-Experiment als auch die RAT in der Lage sind, das stärkere AfD-Signal aufzunehmen, wenn auch nicht ideal. Die Frage, ob die Methoden auch potentielle AfD-Wähler und Wählerinnen miteinbeziehen können, die sich nicht offen als solche zu erkennen geben, bleibt aber zunächst offen. Wir beantworten die Frage mit zwei verschiedenen Methoden, die beide auf der angegebenen ideologischen Nähe zur AfD basieren, denn die Umfrageteilnehmer und -teilnehmerinnen wurden in derselben GIP-Welle gebeten, sich und die sieben größten Parteien auf einer Links-Rechts-Achse zu verordnen.[15]

Für die erste Validierung berechnen wir die wahrgenommene Distanz zwischen jedem/jeder Befragten und allen Parteien und bezeichnen jede Person als der AfD *nahestehend*, für die die AfD unter den zwei nächststehenden Parteien ist. Wir nutzen dies als alternativen Indikator für das AfD-Wählerpotential. Wir identifizieren unter unseren 2597 Befragten 135 Personen, die keine Partei näher an sich selbst positionieren als die AfD (5,2 Prozent). Weitere 156 (6 Prozent) Befragte stehen nur einer anderen Partei näher als der AfD und 141 Personen (5,4 Prozent) sehen sowohl die AfD als auch mindestens eine andere Partei als am nächststehenden an. Insgesamt beziffern wir also das AfD-Wählerpotential aufgrund von ideologischer Nähe auf 16,6 Prozent der Befragten.

Wir treffen die Annahme, dass jede Person, die von uns als der AfD nahestehend identifiziert wurde, auch zum Wählerpotential der AfD gehört. Mit 16,6 Prozent der Befragten ist es deutlich höher als der Wert, den die „Sonntagsfrage“ vorhersagt oder das Wahlergebnis anzeigt. Sowohl die RAT als auch das doppelte Listen-Experiment schließen diesen Wert aber klar in ihre Konfidenzintervalle ein, wie in Abbildung 1 zu sehen ist. Diese Ergebnisse bieten erste Evidenz, dass unsere Schätzungen mithilfe des Listen-Experiments und des RAT-Verfahrens das Wählerpotential der AfD erfassen.

Die zweite Strategie zur Überprüfung, ob das AfD-Wählerpotential valide ermittelt wird, beruht auf den Ergebnissen des Listen-Experiments. Für das Listen-Experiment ist eine Beantwortung der Frage, ob potenzielle AfD-Wähler und Wählerinnen erfasst werden, mit den vorliegenden GIP-Daten möglich. Hierzu nehmen wir an, dass Personen, die sich politisch in der Nähe der AfD verorten, eine höhere Wahrscheinlichkeit haben, AfD zu wählen und deswegen im Listen-Experiment AfD-Wählen als Punkt, den sie wahrscheinlich tun werden, häufiger mitzählen. Des Weiteren nehmen wir an, dass die zwei Listen des doppelten Listen-Experiments (siehe Tabelle 1) hinreichend stark miteinander korrelieren, sodass eine Veränderung im Antwortverhalten lediglich auf das Hinzufügen des Treatments in der Treatmentliste zurückzuführen ist.[16] Vergleicht man nun also die Antworten, die Befragte auf die Treatmentliste und die Kontrollliste abgeben, so sollten Befragte, die der AfD nahestehen, eine höhere Anzahl für die Treatmentliste als für die Kontrollliste angeben. Wir beschränken uns bei unserer Analyse dazu auf Befragte, die in der „Sonntagsfrage“ *nicht* angeben, AfD wählen zu wollen. Sollte der Zusammenhang zwischen wahrgenommener Nähe zur AfD und einer höheren Anzahl affirmativer Antworten für die Treatmentliste dennoch bestehen, so ist dies starke Evidenz, dass das Listenexperiment potenzielle AfD-Wähler messen kann, die die „Sonntagsfrage“ nicht erfasst.

[15] CDU und CSU wurden dabei als CDU/CSU abgefragt.

[16] Entsprechende Tests zeigen, dass der Korrelationskoeffizient 0,25 beträgt und statistisch hochsignifikant ist ($p<0,001$).

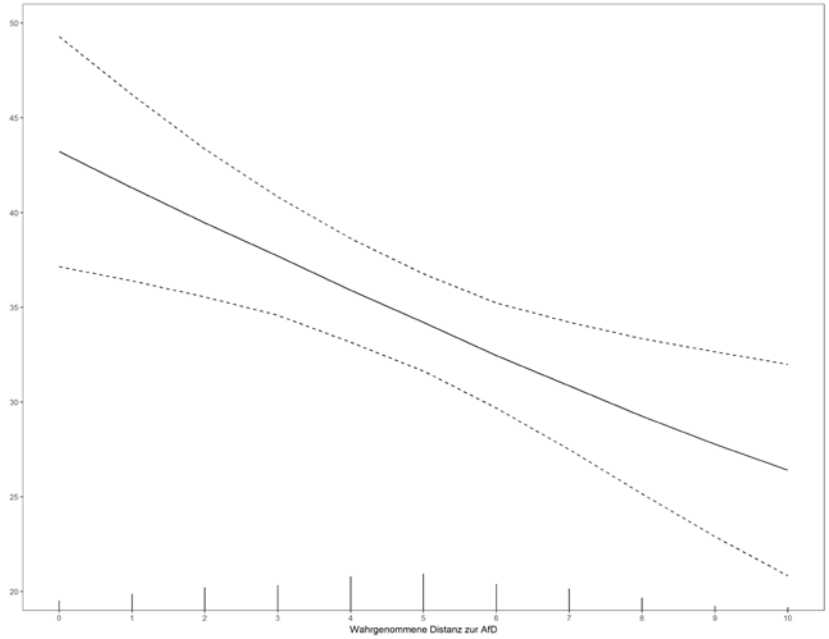
Wir berechnen dazu eine Logit-Regression, die zeigt, wie die Wahrscheinlichkeit, mehr affirmative Antworten in der Treatmentliste als in der Kontrollliste zu geben, von der wahrgenommenen Distanz zur AfD abhängt. Die abhängige Variable der Analyse ist ein binärer Indikator, der den Wert 1 annimmt, wenn der Befragte oder die Befragte angibt, mehr Punkte aus der Treatmentliste als aus der Kontrollliste tun zu werden. Andernfalls nimmt der Indikator den Wert 0 an.[17] Dabei erwarten wir, dass die Wahrscheinlichkeit, mehr affirmative Antworten in der Treatmentliste als in der Kontrollliste zu geben, dann sinkt, wenn der wahrgenommene Abstand zur AfD steigt. Die Ergebnisse befinden sich in Tabelle 2 und Abbildung 3.

Tabelle 2: Validierung des gemessenen AfD-Wählerpotentials basierend auf den Ergebnissen des Listen-Experiments

	Koeffizient (Standardfehler)
Wahrgenommene Distanz zur AfD	-0,08 (0,03)**
Konstante	-0,28 (0,13)*
N	1395

* p<0,05; ** p<0,01

Abbildung 3: Erwartete Wahrscheinlichkeit, im Listenexperiment in der Treatmentliste mehr affirmative Antworten zu geben als in der Kontrollliste (nur Befragte, die in der „Sonntagsfrage“ *nicht* angeben, AfD wählen zu wollen)



Anmerkungen: Die durchgezogene Linie stellt den Punktschätzer dar, die gestrichelten Linien die entsprechenden 95 Prozent-Konfidenzintervalle. Die vertikalen Linien bilden ein Histogramm der Verteilung der wahrgenommenen Distanzen. Quelle: GIP Welle 30.

Wie Abbildung 3 verdeutlicht haben Befragte, die in der „Sonntagsfrage“ keine AfD-Wahlabsicht bekunden und sich selbst sehr nah zur AfD einschätzen, eine über 40 prozentige Wahrscheinlichkeit, in der Treatmentliste mehr affirmative Antworten zu geben als in der Kontrollliste. Steigt die gefühlte Distanz zur AfD, so sinkt diese Wahrscheinlichkeit rasch ab.

[17] Befragte, die nicht beide Fragen beantworten, wurde bereits in den anderen Analysen ausgeschlossen. Wir tun dies auch hier.

Bei einer gefühlten Distanz von 5 auf der 11-Punkte-Skala ist die Wahrscheinlichkeit schon um 9 Prozentpunkte gesunken. Dieser Zusammenhang ist statistisch signifikant und substantiell relevant. Wir sehen in diesem Befund klare Hinweise darauf, dass das Listenexperiment in der Lage ist, potenzielle AfD-Wähler zu erfassen, die von der „Sonntagsfrage“ nicht erfasst würden. Dies ist ein klarer Vorteil des Listenexperiments gegenüber der „Sonntagsfrage“.

5. Schlussfolgerungen

Es wird nicht überraschen, aber am nächsten Sonntag ist üblicherweise keine Bundestagswahl. Trotzdem sind die individuellen Antworten auf die hypothetische Frage nach der Wahlabsicht an jenem Tag zentral für die Wahl- und Einstellungsforschung. Darüber hinaus spielen die aggregierten Ergebnisse der „Sonntagsfrage“ eine bedeutende Rolle für die Parteienforschung und nicht zuletzt für die Medien und dadurch vermittelt auch für die breite Öffentlichkeit.

Die wesentlichen Schwächen und daraus resultierende Einschränkungen dieser einfachen Frage im politikwissenschaftlichen Diskurs sind hinlänglich bekannt. Lediglich adäquate und praktikable Alternativen scheinen bislang noch nicht gefunden worden zu sein. Ziel des Beitrags ist es am Beispiel der Wahlabsicht für die rechtspopulistische Partei AfD zu evaluieren, inwiefern es hinsichtlich dieser Schwächen der „Sonntagsfrage“ alternative Instrumentierungen gibt, die eine valide Messung ermöglichen. Dazu implementierten wir drei neue Instrumente im Rahmen einer repräsentativen Bevölkerungsumfrage des German Internet Panels. Der direkte Vergleich der alternativen Verfahren mit der „Sonntagsfrage“ zeigt, dass alle drei Alternativen den Stimmanteil der AfD als signifikant höher schätzen. Allerdings sind die Konfidenzintervalle der Randomisierten-Antwort-Technik und des Listen-Experiments so groß, dass der praktische Nutzen dieser Verfahren zu Prognosezwecken stark eingeschränkt ist. Die Weisheit-der-Vielen Methode scheint hingegen eine nützliche Ergänzung zur klassischen „Sonntagsfrage“ darzustellen, da sie sowohl den genauesten Punktschätzer als auch das kleinste Konfidenzintervall aufweist. Dies liegt vermutlich daran, dass diese Methode lediglich geringe kognitive Ansprüche an die Befragten stellt und keinen Widerspruch zwischen einer möglichst genauen Schätzung des AfD-Wählerpotentials und der Anonymität der Befragten erzeugt. Da ein solches Instrument außerdem leicht zu implementieren, analysieren und vergleichsweise günstig ist, unterstützen die Ergebnisse in dieser Studie die Empfehlung von Graefe (2014) und Murr (2016), dass die akademische, aber auch die kommerzielle Meinungsforschung die Weisheit-der-Vielen-Methode öfter nutzen sollten. Sie scheint insbesondere mit Blick auf populistische Parteien gut geeignet zu sein, die traditionelle Berichterstattung hinsichtlich der „Sonntagsfrage“ sinnvoll zu ergänzen.

Im Unterschied zu Johan et al. (2016) sind wir nicht nur an relativen Unterschieden in der Potentialmessung der möglichen AfD-Wählerschaft zwischen mehreren Methoden interessiert. Wir zeigen explizit, dass sowohl unsere Schätzungen, basierend auf dem doppelten Listen-Experiment sowie der RAT, statistisch nicht verschieden vom amtlichen Endergebnis sind. Das ist zumindest eine Form der externen Validierung, die wir vornehmen können, weil die Größe des tatsächlichen Wählerpotentials für die AfD nicht bekannt ist. Zudem zeigen alle drei hier untersuchten Methoden, dass sie in der Lage sind, einen höheren Anteil an sozial unerwünschten Antworten zu Tage zu fördern als die traditionelle „Sonntagsfrage“ selbst.

Weitergehende Robustheitstests zeigen, dass sich die relativen Unterschiede zwischen den Methoden auch nicht verändern, wenn man gegebenenfalls übereilte Antworten aus der Ana-

lyse ausschließt. Da bekanntlich Online-Umfragen weniger anfällig für Verzerrungen aufgrund von sozialer Erwünschtheit sind (Kreuter et al. 2008), ist unsere Studie ein harter Test, weil den Befragten aufgrund des Online-Modus bereits mehr Privatheit gewährt wird als beispielsweise in Telefonumfragen oder gar persönlichen Interviews. Alle drei Methoden erlauben es allerdings nicht, auf der Individualebene, etwa für einen bestimmten Wählertyp, vorherzusagen, wie hoch die Neigung ist, AfD zu wählen. Das ist offenbar der Preis, den man zahlen muss, um Befragten mehr Privatheit bzw. Anonymität zu gewähren und dadurch dem Problem der sozialen Erwünschtheit zu begegnen.

Nichtsdestotrotz erweist sich die Strategie, Störgeräusche einzubauen, um den Befragten mehr Anonymität zu gewähren und letztlich so ein valides Signal herauszufiltern, als vielversprechend zur Messung der AfD-Unterstützung. Unsere empirischen Tests legen nahe, dass die Unterschiede zwischen den verschiedenen Messmethoden wesentlich dadurch beeinflusst werden, dass sich potentielle AfD-Wähler nicht als solche in der „Sonntagsfrage“ zu erkennen geben.

Insgesamt konnten wir zeigen, dass das AfD-Wählerpotential bereits im Juli bedeutend größer war als auf Basis der „Sonntagsfrage“, aber auch auf Basis anderer zu diesem Zeitpunkt erhobener Umfragen zu erwarten gewesen wäre. Gleichzeitig gilt auch, dass nicht jeder/jede, der/die sich vorstellen kann die AfD zu wählen, das auch am Wahltag tatsächlich tut. Das Wählerpotential einer Partei ist notwendigerweise größer als ihre tatsächliche Wählerschaft.

Wir führten eine weitere Validierung mittels Befragter, die sich bei der Wahlabsichtsfrage zur AfD Wahl bekennen, durch. Unter der Annahme, dass sich diese Befragten nicht nur auf die direkte Frage nach der Wahlabsicht zur AfD bekennen, sondern auch Teil des Wählerpotentials der AfD sein müssen, das mittels des doppelten Listen-Experiments sowie der RAT bestimmt wurde, sollten alle AfD-Wähler und Wählerinnen auch Teil des geschätzten Wählerpotentials sein. Wir konnten diese Implikation nur beim Listen-Experiment finden. Die RAT schätzt ein systematisch niedrigeres Wählerpotential als durch die Antworten auf die direkte Frage eigentlich zu erwarten gewesen wäre. Aus der Einstellungsforschung wissen wir, dass Messinstrumente, die auf nur einer Frage beruhen – wie eben die Wahlabsichtsfrage – Messfehler aufweisen werden, die die Annahme dieses Validierungstestes untergraben (z.B. Behnke et al. 2006). Daher ist es vorschnell den Stab über der RAT zu brechen. Mehr Forschung ist nötig, die über die „Sonntagsfrage“ hinausgeht, um auch die Validität der Wahlabsichtsfrage selbst nicht nur anzunehmen, sondern auch tatsächlich zu testen. Die Validierung der Wahlabsichtsfrage ist ein lohnenswertes Forschungsprogramm, das die Wahl- und Einstellungsforschung bisher nicht angegangen ist. Auch die Autoren dieser Studie nehmen in ihren eigenen Arbeiten bisher an, dass die Antworten zur Wahlabsichtsfrage fehlerfrei gemessen werden.

Die Ergebnisse unsere Studie haben schließlich noch weitere wichtige Implikationen für die zukünftige Arbeit der Wahl- und Parteienforschung. Die aggregierten Ergebnisse der Wahlabsichtsfrage sind zentral für die Parteienforschung. Sie werden in vielen Arbeiten als abhängige Variable benutzt um zu zeigen, welchen Effekt die Veränderung einer Parteienstrategie oder programmatischen Positionierung am Wahltag hat. Viele Arbeiten nutzen sie aber auch als unabhängige Variable, um etwa die Änderungen in Parteien- oder Koalitionsverhalten zu erklären.

Unsere Methoden können auch helfen, die Faktoren zu bestimmen, die eine Partei als mehr oder weniger sozial erwünscht erscheinen lassen. Ausgangpunkt ist dabei, dass deutliche Abweichungen zwischen Umfrageergebnissen und Wahlverhalten auf soziale Erwünschtheit

schließen lassen, also zum Beispiel darauf, dass es viele Menschen als sozial unerwünscht ansehen, AfD zu wählen und ihre AfD-Wahlabsicht deshalb in Umfragen nicht preisgeben. Da Wahlen aber nur in sehr weiten Abständen zu einander stattfinden, ist es nahezu unmöglich herauszukristallisieren, welches von vielen Ereignissen, die zwischen zwei Wahlen stattfinden, dazu führen, dass die soziale Akzeptanz einer Partei steigt oder sinkt. Mit Hilfe der hier vorgestellten Methoden ist es aber möglich, dass AfD-Wählerpotenzial auch zwischen Wahlen genauer zu bestimmen. Vergleicht man das mit den hier vorgestellten Methoden ermittelte Potenzial mit dem Anteil an Befragten, die sich offen zu einer AfD-Wahlabsicht bekennen („Sonntagsfrage“), erhält man ebenfalls eine Messung der sozialen Erwünschtheit der AfD. Misst man in dieser Form die soziale Erwünschtheit der AfD etwa kurz bevor und kurz nachdem sie wichtige Personal- oder Programmentscheidungen trifft, so kann man den Einfluss solcher Entscheidungen auf ihre soziale Erwünschtheit messen.

Schlussendlich sind die Messmethoden, die wir hier vorstellten und testeten geeignet, die politikwissenschaftliche Forschung voranzutreiben und uns zu weiteren Erkenntnisse zu verhelfen.

Literatur

Adams, James, Michael Clark, Lawrence Ezrow, und Garrett Glasgow. 2004. Understanding Change and Stability in Party Ideologies: Do Parties Respond to Public Opinion or to Past Election Results? *British Journal of Political Science* 34(4), 589–610. DOI: https://doi.org/10.1017/S0007123404000201.

Aichholzer, Julian, Sylvia Kritzinger, Markus Wagner, und Eva Zeglovits. 2014. How has Radical Right Support Transformed Established Political Conflicts? The Case of Austria. *West European Politics* 37(1), 113–137. DOI: https://doi.org/10.1080/01402382.2013.814956.

Behnke, Joachim, Nina Baur, und Nathalie Behnke. 2006. *Empirische Methoden der Politikwissenschaft*, Paderborn: Schöningh.

Bergmann, Knut, und Matthias Diermeier. 2017. *Die AfD: Eine unterschätzte Partei. Soziale Erwünschtheit als Erklärung für fehlerhafte Prognosen,*.

Blom, Annelies G., Jessica M. E. Herzing, Carina Cornesse, Joseph W. Sakshaug, et al. 2016. Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review* 35(4), 498–520. DOI: https://doi.org/10.1177/0894439316651584.

Boland, Philip J. 1989. Majority Systems and the Condorcet Jury Theorem. *Journal of the Royal Statistical Society. Series D (The Statistician)* 38(3), 181–189. DOI: https://doi.org/10.2307/2348873.

Bräuninger, Thomas. 2009. Responsivität und strategische Adaption im Parteienwettbewerb in den deutschen Bundesländern. In *Parteienwettbewerb, Wählerverhalten und Koalitionsbildung : Festschrift zum 70. Geburtstag von Franz Urban Pappi*, Hrsg. Christian Henning, Eric Linhart, und Susumu Shikano, 27–46. Baden-Baden: Nomos.

Bräuninger, Thomas, und Marc Debus. 2012. *Parteienwettbewerb in den deutschen Bundesländern*, Wiesbaden: VS Verlag für Sozialwissenschaften.

Condorcet, Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris: De I'Imprierie Royale.

Coutts, Elisabeth, und Ben Jann. 2011. Sensitive questions in online surveys: Experimental results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research* 40(1), 169–193. DOI: https://doi.org/10.1177/0049124110390768.

Crowne, Douglas P., und David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology* 24(4), 349–354. DOI: https://doi.org/10.1037/h0047358.

Diekmann, Andreas. 2012. Making Use of “Benford’s Law” for the Randomized Response Technique. *Sociological Methods & Research* 41(2), 325–334. DOI: https://doi.org/10.1177/0049124112452525.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, et al. 1991. The Item-Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application. In *Measurement Errors in Surveys*, Hrsg. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, et al., 185–210. Hoboken, NJ, USA: John Wiley & Sons Inc.

Evans, Jocelyn, und Gilles Ivaldi. 2010. Comparing forecast models of Radical Right voting in four European countries (1973-2008). *International Journal of Forecasting* 26(1), 82–97. DOI: https://doi.org/10.1016/J.IJFORECAST.2009.04.001.

Galton, Francis. 1907. Vox populi. *Nature* 75, 450–451.

Glynn, Adam N. 2013. What can we learn with statistical truth serum? *Public Opinion Quarterly* 77(S1), 159–172. DOI: https://doi.org/10.1093/poq/nfs070.

Graefe, Andreas. 2014. Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly* 78(S1), 204–232. DOI: https://doi.org/10.1093/poq/nfu008.

Grofman, Bernard, Guillermo Owen, und Scott L. Feld. 1983. Thirteen theorems in search of the truth. *Theory and Decision* 15(3), 261–278. DOI: https://doi.org/10.1007/BF00125672.

Groß, Jochen. 2010. *Die Prognose von Wahlergebnissen. Ansätze und empirische Leistungsfähigkeit*, Wiesbaden: VS Verlag für Sozialwissenschaften.

Höglinger, Marc, und Andreas Diekmann. 2017. Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis* 25(1), 131–137. DOI: https://doi.org/10.1017/pan.2016.5.

Höglinger, Marc, Ben Jann, und Andreas Diekmann. 2016. Sensitive Questions in Online Surveys: An Experimental Comparison of the Randomized Response Technique and the Crosswise Model. *Survey Research Methods* 10(3), 171–187. DOI: https://doi.org/10.18148/srm/2016.v10i3.6703.

Jann, Ben, Julia Jerke, und Ivar Krumpal. 2012. Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly* 76(1), 32–49. DOI: https://doi.org/10.1093/poq/nfr036.

Johann, David, Kathrin Thomas, Thorsten Faas, und Sebastian Fietkau. 2016. Alternative Messverfahren rechtspopulistischen Wählens im Vergleich: Empirische Erkenntnisse aus Deutschland und Österreich. In *Wahlen und Wähler: Analysen aus Anlass der Bundestagswahl 2013*, Hrsg. Harald Schoen, und Bernhard Weßels, 447–470. Wiesbaden: Springer Fachmedien Wiesbaden.

Kollman, Ken, John H. Miller, und Scott E. Page. 1992. Adaptive Parties in Spatial Elections. *American Political Science Review* 86(4), 929–937. DOI: https://doi.org/10.2307/1964345.

Kreuter, Frauke, Stanley Presser, und Roger Tourangeau. 2008. Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly* 72(5), 847–865. DOI: https://doi.org/10.1093/poq/nfn063.

Krosnick, Jon A. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5(3), 213–236. DOI: https://doi.org/10.1002/acp.2350050305.

Krosnick, Jon A. 1999. Survey Research. *Annual Review of Psychology* 50(1), 537–567. DOI: https://doi.org/10.1146/annurev.psych.50.1.537.

Krumpal, Ivar. 2012. Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research* 41(6), 1387–1403. DOI: https://doi.org/10.1016/j.ssresearch.2012.05.015.

Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, et al. 1997. Racial Prejudice and Attitudes Toward Affrimative Action. *American Journal of Political Science* 41, 402–419.

Lehmbruch, Gerhard. 2000. *Parteienwettbewerb im Bundesstaat: Regelsysteme und Spannungslagen im politischen System der Bundesrepublik Deutschland*, Wiesbaden: VS Verlag für Sozialwissenschaften.

Lewandowsky, Marcel. 2015. Eine rechtspopulistische Protestpartei? Die AfD in der öffentlichen und politikwissenschaftlichen Debatte. *Zeitschrift für Politikwissenschaft* 25(1), 119–134. DOI: https://doi.org/10.5771/1430-6387-2015-1-119.

Lewandowsky, Marcel, Heiko Giebler, und Aiko Wagner. 2016. Rechtspopulismus in Deutschland. Eine empirische Einordnung der Parteien zur Bundestagswahl 2013 unter besonderer Berücksichtigung der AfD. *Politische Vierteljahresschrift* 57(2), 247–275. DOI: https://doi.org/10.5771/0032-3470-2016-2-247.

Lewis-Beck, Michael S. 2005. Election Forecasting: Principles and Practice. *The British Journal of Politics and International Relations* 7(2), 145–164. DOI: https://doi.org/10.1111/j.1467-856X.2005.00178.x.

List, Christian, und Robert E. Goodin. 2001. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy* 9(3), 277–306. DOI: https://doi.org/10.1111/1467-9760.00128.

Munzert, Simon, Lukas Stötzer, Thomas Gschwend, Marcel Neunhoeffer, et al. 2017. Zweitstimme.org. Ein strukturell-dynamisches Vorhersagemodell für Bundestagswahlen. *Politische Vierteljahresschrift* 58(3), 418–441. DOI: https://doi.org/10.5771/0032-3470-2017-3-418.

Murr, Andreas E. 2017. Wisdom of Crowds. In *Handbook of Political Behavior*, Hrsg. Kai Arzheimer, Jocelyn Evans, und Michael S. Lewis-Beck, 835–860. Los Angeles: Sage.

Murr, Andreas E. 2016. The wisdom of crowds: What do citizens forecast for the 2015 British General Election? *Electoral Studies* 41, 283–288. DOI: https://doi.org/10.1016/j.electstud.2015.11.018.

Phillips, Derek L., und Kevin J. Clancy. 1972. Some Effects of „Social Desirability“ in Survey Studies. *American Journal of Sociology* 77(5), 921. DOI: https://doi.org/10.1086/225231.

Rosenfeld, Bryn, Kosuke Imai, und Jacob N. Shapiro. 2016. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science* 60(3), 783–802. DOI: https://doi.org/10.1111/ajps.12205.

Rosenfelder, Joel. 2017. Die Programmatik der AfD: Inwiefern hat sie sich von einer primär euroskeptischen zu einer rechtspopulistischen Partei entwickelt? *Zeitschrift für Parlamentsfragen* 48(1), 123–140. DOI: https://doi.org/10.5771/0340-1758-2017-1-123.

Roth, Dieter, und Andreas M. Wüst. 2015. Missgeschick oder Trend? Zur Prognosetauglichkeit von Wahlumfragen. *Zeitschrift für Staats- und Europawissenschaften* 13(2), 298–330. DOI: https://doi.org/10.5771/1610-7780-2015-2-298.

Schmitt-Beck, Rüdiger, Hans Rattinger, Sigrid Roßteutscher, Berhard Weßels, et al. 2014. *Zwischen Fragmentierung und Konzentration: die Bundestagswahl 2013*, Baden-Baden: Nomos.

Schnell, Rainer, und Marcel Noack. 2014. The Accuracy of Pre-Election Polling of German General Elections. *Methods, Data, Analyses* 8(1), 5–24. DOI: https://doi.org/10.12758/mda.2014.001.

Wüst, Andreas M. 2003. Stimmung, Projektion, Prognose? In *Politibarometer*, Hrsg. Andreas M. Wüst, 83–107. Wiesbaden: VS-Verlag.

Yu, Jun-Wu, Guo-Liang Tian, und Man-Lai Tang. 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika* 67(3), 251–263. DOI: https://doi.org/10.1007/s00184-007-0131-x.

Anhang

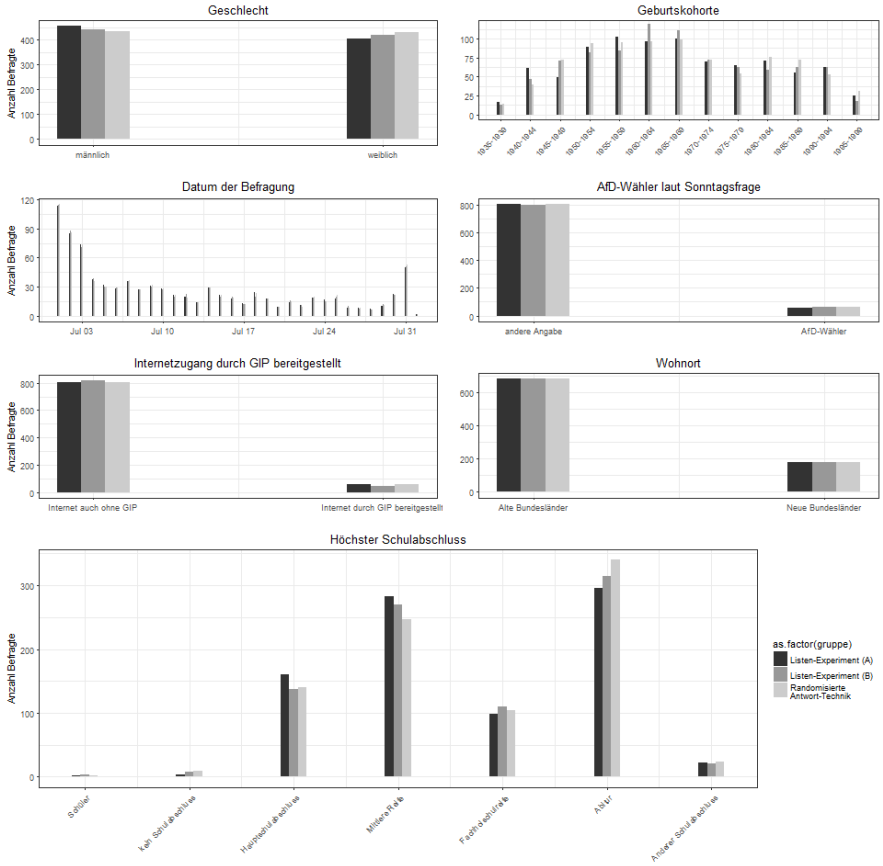
A1. Test der Randomisierung

Wir teilen Befragte zufällig in drei Gruppen ein, die jeweils einen bestimmen Fragentyp gestellt bekommen. Ein zufällig ausgewähltes Drittel der Befragten wurde der RAT-Befragung zugeordnet, das zweite zufällig ausgewählte Drittel erhielt die Listen-Experiment-Fragen mit Liste A als Treatment und Liste B als Kontrollliste und das letzte Drittel der Befragten erhielt ebenfalls die Listen-Experiment-Fragen, jedoch mit Liste B als Treatment und Liste A als Kontrollliste. Die zufällige Einteilung soll erreichen, dass die drei Gruppen in allen wesentlichen Merkmalen gleich sind. In diesem Anhang zeigen wir, dass dies zwar nicht perfekt, aber doch in beachtlicher Weise gelingt. Abbildung A1 zeigt die Anzahl der Befragten, die bestimmten Charakteristika zugeordnet werden können. Diese sind das (1) Geschlecht der Befragten, (2) eine etwa fünfjährige Jahresspanne, in der die Befragten geboren wurden, (3) das Datum, an dem die Befragten die Umfrage ausfüllten, (4) ob sie in der „Sonntagsfrage“ eine AfD-Wahlabsicht äußern, (5) ob sie in den alten oder neuen Bundesländern wohnen, (6) welcher ihr höchster Schulabschluss ist und (7) ob sie auch ohne das GIP über einen Internetzugang verfügen.

Die letzte Eigenschaft bedarf einer kurzen Erklärung. Bei der Rekrutierung der GIP-Teilnehmer und Teilnehmerinnen wurde die gesamte deutsche Bevölkerung zwischen 16 und 75 Jahren zugrunde gelegt und dies unabhängig davon, ob die ausgewählten Personen über einen Internetzugang oder die Computerkenntnisse verfügen, um an einer Online-Befragung teilzunehmen. Befragten, die über mindestens einen dieser Faktoren nicht verfügen, wurden sowohl ein für Computer-Anfänger entwickelter Computer sowie ein Internetzugang bereitgestellt (Blom et al. 2016). Wir vergleichen deshalb zwischen den Gruppen, wie viele Personen vom GIP-Team mit einem Internetanschluss oder Computer ausgestattet wurden.

Die verschiedenen grauen Säulen in Abbildung A1 stellen die verschiedenen Gruppen dar. Sind diese Säulen annähernd gleichhoch, so gleichen sich die Gruppen in der dargestellten Eigenschaft. Auf den ersten Blick ist sichtbar, dass die Gruppen sich tatsächlich stark gleichen (siehe etwa Geschlecht, AfD-Wähler/Wählerin laut „Sonntagsfrage“, Internetzugang durch GIP bereitgestellt und Wohnort). Die Eigenschaften mit vielen Merkmalsausprägungen (Datum der Befragung, Geburtenkohorte und höchster Schulabschluss) weichen zwar in Einzelfällen voneinander ab, würden sich aber annähernd wieder ausgleichen, wenn benachbarte Kategorien zusammengefasst würden. Da die Einteilung in die Kategorien zu einem gewissen Grad willkürlich ist (zum Beispiel die Einteilung der Geburtskohorten), sind wir überzeugt, dass die vereinzelt sichtbaren Abweichungen unsere Ergebnisse nicht wesentlich beeinflussen.

Abbildung A1: Verteilung zentraler Charakteristika unter den Befragungsgruppen



Anmerkungen: Alle Grafiken sind Histogramme. Listen-Experiment (A) = Liste A enthält fünf Aussagen, Listen-Experiment (B) = Liste B enthält fünf Aussagen. Quelle: GIP Welle 30.

A2. Verteilung der letzten Ziffer von Hausnummern in Deutschland

Die Verteilung der ersten Ziffern der Hausnummern aller GIP Befragten haben wir ausgewertet, um die Annahme der Bendforschen Verteilung zu überprüfen. Das ist ein zulässiger Test, da es sich bei den GIP Befragten um eine repräsentative Stichprobe der Bevölkerung handelt. Abbildung A2 zeigt, dass unsere Verteilungsannahme gerechtfertigt ist. Circa 70 % der Hausnummern der Befragten beginnen mit Ziffern zwischen 1 und 4.

Abbildung A2: Verteilung der letzten Ziffer der Hausnummern der GIP-Befragten